Lectures on Nonparametric Bayesian Statistics

Notes for the course by Bas Kleijn, Aad van der Vaart, Harry van Zanten
(Text partly extracted from a forthcoming book by S. Ghosal and A. van der Vaart)
version 4-12-2012

UNDER CONSTRUCTION

1

Introduction

Why adopt the nonparametric Bayesian approach for inference? The answer lies in the simultaneous preference for nonparametric modeling and desire to follow a Bayesian procedure. Nonparametric (and semiparametric) models can avoid the arbitrary and possibly unverifiable assumptions inherent in parametric models. Bayesian procedures may be desired for the conceptual simplicity of the Bayesian paradigm, easy interpretability of Bayesian quantities or philosopohical reasons.

1.1 Motivation

Bayesian nonparametrics is the study of Bayesian inference methods for nonparametric and semiparametric models. In the Bayesian nonparametric paradigm a prior distribution is assigned to all unknown quantities (parameters) involved in the modeling, whether finite or infinite dimensional. Inference is made from the "posterior distribution", the conditional distribution of all parameters given the data. A *model* completely specifies the conditional distribution of all observable given all unobserved quantities, or parameters, while a *prior distribution* specifies the distribution of all unobservables. From this point of view, random effects and latent variables also qualify as parameters, and distributions of these quantities, often considered as part of the model itself from the classical point of view, are considered part of the prior. The *posterior distribution* involves an inversion of the order of conditioning. Existence of a regular version of the posterior is guaranteed under mild conditions on the relevant spaces (see Section 2).

1.1.1 Classical versus Bayesian nonparametrics

Nonparametric and semiparametric statistical models are increasingly replacing parametric models, to overcome the latter's inflexibility to address a wide variety of data. A nonparametric or semiparametric model involves at least one infinite-dimensional parameter and hence may also be referred to as an "infinite-dimensional model". Indeed, the nomenclature "nonparametric" is misleading in that it gives the impression that there is no parameter in the model, while in reality there are infinitely many unknown quantities. However, the term nonparametric is so popular that it makes little sense not to use it. The infinite-dimensional parameter is usually a function or measure. In a canonical example of nonparametric model the data are a random sample from a completely unknown distribution P. More generally, functions of interest include the cumulative distribution function, density function, regression function, hazard rate, transition density of a Markov process, spectral density of a time series, response probability of a binary variable as a function of covariates, false discovery rate as a function of nominal level in multiple testing, receiver operating characteristic function between two distributions. Non-Bayesian methods for the estimation of many of these functions have been well developed, partly due to the availability of simple, natural and well respected estimators (e.g. the empirical distribution), and partly driven by the greater attention historically given to the classical approach. Bayesian estimation methods for nonparametric problems started receiving attention in the last three decades.

Different people may like Bayesian methods for different reasons. To some the appeal is mainly philosophical. Certain *axioms of rational behavior* naturally lead to the conclusion that one ought to follow a Bayesian approach not to irrational (see Bernardo and Smith (1995)). Although the axioms themselves can be questioned, there is a wide-spread impression that Bayesian methods are perhaps logically more consistent than non-Bayesian methods, particularly compared to those which involve integration over the sample space, that is, methods which "bother about samples that could have but have not realized". Others justify the Bayesian paradigm by appealing to exchangeability and de Finetti's (1937) theorem. This celebrated theorem concludes the existence of a "random parameter" instead of a "fixed parameter" based on a "concrete" set of observations and a relatively weak assumption on distributional invariance (see Schervish (1996)). However, this requires subjectively specifying a prior, which is considered as difficult. Decision theoretists may like to be Bayesians, because of the complete class theorem, which asserts that for any procedure there is a better Bayes procedure, and only the latter procedures are admissible, or essentially so (see Ferguson (1967)). While this could have been a strong reason for a frequentist to take the Bayesian route, there are difficulties. First, the complete class theorem is known to hold only when the parameter space is compact (and the loss function is convex), and secondly, it is not clear which prior to choose from the class of all the priors. People who believe in asymptotic theory could find Bayesian methods attractive because their large sample optimality (in parametric problems). However, this argument is weak, because many non-Bayesian procedures (most notably, the maximum likelihood estimator (MLE)) are also asymptotically optimal.

The specification of a prior distribution may be challenging. Keeping this aside, the Bayesian approach is extremely straightforward, in principle — the full inference is based on the posterior distribution only. All inference tools are produced by one stroke — one need not start afresh when the focus of attention changes from one quantity to another. In particular, the same analysis produces an estimate as well as an assessment of its accuracy (in terms of variability or a probable interval of the location of the parameter value). The Bayesian approach produces a "real probability" on the unknown parameter as a quantification of the uncertainty about its value. This comes extremely handy, for instance, in the construction of intervals and tests. On the other hand, sampling variability is sometimes difficult to visualize given data, and often leads to logical problems. When the parameter space is restricted, many of the non-Bayesian methods, including the MLE can be unnatural. The Bayesian approach also eliminates the problem of nuisance parameter by simply integrating them out, while classical procedures will often have to find ingenious ways to tackle them separately for each inference problem. Prediction problems, which are often considered to be the primary objective of statistical analysis, are solved most naturally if one follows the Bayesian approach.

However, these conveniences come at a price. The Bayesian principle is also restricting in nature, allowing no freedom beyond the choice of the prior. This can put Bayesian methods at a disadvantage vis-a-vis non-Bayesian methods, particularly when the performance is evaluated by frequentist principles. For instance, even if only a part of the unknown parameter is of interest, a Bayesian still has to specify a prior on the whole parameter, compute the posterior, and integrate out the irrelevant part, whereas a classical procedure may be able to target the part of interest only. Another problem is that no corrective measure is allowed in a Bayesian framework once the prior has been specified. In contrast, the MLE is known to be non-existent or inconsistent in many infinite-dimensional problems such as density estimation, but one can modify it by penalization, sieves (see Grenander (1981)), partial likelihood (Cox (1972)) or other devices. In contrast, the Bayesian principle does not allow an honest Bayesian to change the likelihood, change the prior by looking at the data, or even change the prior with increasing sample size.

1.1.2 Parametric versus nonparametric Bayes

Parametric models make restrictive assumptions about the data generating mechanism, which may cause serious bias in inference. In the Bayesian framework a parametric model assumption can be viewed as an extremely strong prior opinion. Indeed, a parametric model specification $X|\,\theta \sim p_\theta$, for $\theta \in \Theta \subset \mathbb{R}^d$, with a prior $\theta \sim \pi$, may be considered within a nonparametric Bayesian framework as $X|\,p \sim p$, for $p \in \mathcal{P}$ with $\mathcal{P}$ a large set of densities equipped with a prior $p \sim \Pi$ with the property that $\Pi\big(\{p_\theta\colon \theta \in \Theta\}\big) = 1$. Thus parametric modelling is equivalent to insisting on a prior that assigns probability one to a thin subset of all densities. This is a very strong prior opinion indeed.

To some extent the nonparametric Bayesian approach also solves the problem of partial specification. Often a model is specified incompletely, without describing every detail of the data generating mechanism. A familiar example is the Gauss-Markov setup of a linear model, where errors are assumed uncorrelated, mean zero variables with constant variance, but no further distributional assumptions are imposed. Lacking a likelihood, a parametric Bayesian approach cannot proceed further. However, a nonparametric Bayes approach can put a prior on the space of densities with mean zero and consider that as a prior on the error distribution. More generally, incomplete model assumptions may be complemented by general assumptions involving infinite-dimensional parameters in order to build a complete model, which a nonparametric Bayesian approach can equip with infinite-dimensional priors.

1.2 Challenges of Bayesian nonparametrics

This section describes some conceptual and practical difficulties that arise in Bayesian nonparametrics, along with possible remedies.

1.2.1 Prior construction

A Bayesian analysis cannot proceed without a prior distribution on all parameters. A prior ideally expresses a quantification of pre-experiment and subjective knowledge. A prior on a function requires knowing many aspects of the function, including infinitesimal details, and the ability to quantify the information in the form of a probability measure. This poses an apparent conceptual contradiction: a nonparametric Bayesian approach is pursued to minimize restrictive parametric assumptions, but at the same time requires specification of the minute details of a prior on an infinite-dimensional parameter.

There seems to exist overall agreement that subjective specification of a prior cannot be expected in complex statistical problems. Instead inference must be based on an *objective prior*. This is vaguely understood as a prior that is proposed by some automatic mechanism that is not in favor of any particular parameter values, and has low information content compared to the data. Instead of "objective prior" we also use the phrase *default prior*.

Some of the earliest statistical analyses in history used the idea of *inverse probability*, which is nothing but a default Bayesian analysis with respect to a uniform prior. Uniform priors were strongly criticised for lacking invariance, which led to temporary scrapping of the idea, but later more invariance-friendly methods such as *reference analysis* or *probability matching* emerged. However, most of these ideas are restricted to finite-dimensional parametric problems.

An objective prior should be automatically constructed using a default mechanism. It need not be non-informative, but should be spread all over the parameter space. Some key hyperparameters regulating the prior may be chosen by the user, whereas other details must be constructed by the default mechanism. Unlike in parametric situations, where non-informative priors are often improper, default priors considered in nonparametric Bayesian inference are almost invariably proper. Large support of the prior means that the prior is not too concentrated in some particular region. This generally makes that the information contained in the prior is subdued gradually by the data if the sample size increases, so that eventually the data override the prior.

The next chapters discuss various methods of prior construction for various problems of interest. Although a default prior is not unique in any sense, it is expected that over the years, based on theoretical results and practical experience, a handful of suitable priors will be short-listed and catalogued for consensus use in each inference problem.

1.2.2 Computation

The property of conjugacy played an important role in parametric Bayesian analysis, as it enabled the derivation of posterior distributions, in a time that computing resources were lacking. Later sampling based methods such as the Metropolis-Hastings algorithm and Gibbs sampling gave Bayesian analysis a tremendous boost. Without modern computing nonparametric Bayesian analysis would hardly be practical.

However, we cannot directly simulate from the posterior distribution of a function unless it is parameterized by finitely many parameters. We must break up the function of interest into more elementary finite-dimensional objects, and simulate from their posterior distribution. For this reason the structure of the prior is important. Useful structure may come from conjugacy or approximation. Often a computational method combines analytic derivation and Markov chain Monte-Carlo (MCMC) algorithms, and is based on really innovative ideas. For instance, density estimation with a Dirichlet mixture prior uses an equivalent hierarchical mixture model involving a latent variable for each observation, and integrates out the infinite-dimensional parameter given the latent variables. Thus the problem of infinite dimension has been reduced to one of dimension smaller than the sample size. In another instance, in a binary response model and a Gaussian process prior, introduction of normal latent variables brings in conjugacy.

1.2.3 Asymptotic behavior

Putting a prior on a large parameter space makes it easy to be grossly wrong. Therefore studying robustness is important in Bayesian nonparametrics. Bayesian robustness means that the choice of the prior does not influence the posterior distribution too much. This is difficult to study in a general framework. A more manageable task is the study of asymptotic properties, as the information in the data increases indefinitely. For example, "posterior consistency" may be considered an asymptotic form of robustness. Loosely speaking posterior consistency means that the posterior probability eventually concentrates in a (any) small neighborhood of the actual value of the parameter. This is a weak property, shared by many prior distributions. Finer properties, such as the rate of convergence or a (functional) limit theorem, give more insight in the performance of different priors.

The study of asymptotic properties is more complex in the nonparametric than in parametric context. In the parametric setting good properties are quaranteed under mild conditions, such as the true value of the parameter being in the support of the prior (provided that the model satisfies some basic regularity conditions). In the infinite-dimensional context merely having the true value in the topological support of the prior is not sufficient. Consistency may fail for very natural priors satisfying the support condition, meaning that even an infinite amount of data may not overcome the pull of a prior in a wrong direction. Consistent priors may amongst themselves differ strongly in accuracy, depending on their fine details. For instance, the rate of contraction of the posterior to the true parameter may strongly depend on the prior. Unlike in the parametric setting many priors do not "wash out" as the information in the data increases indefinitely.

Thus it makes sense to impose posterior consistency and a good rate of contraction as requirements on a "default prior". Several chapters are devoted to the study of asymptotic behavior of the posterior distribution and other related quantities. Attractive is also to combine priors hierarchically into an overall prior, so as to make the posterior "adapt" to a large class of underlying true parameters.

2

Priors, posteriors and Bayes's rule

In the Bayesian framework the data X follows a distribution determined by a parameter θ, which is itself considered to be generated from a *prior distribution* Π. The corresponding *posterior distribution* is the conditional distribution of θ given X. This framework is identical in parametric and nonparametric Bayesian statistics, the only difference being the dimensionality of the parameter. Because the proper definitions of priors and (conditional) distributions require (more) care in the nonparametric case, we review the basics of conditioning and Bayes's rule in this section.

The parameter set Θ is equipped with a σ-field $\mathscr{B}$, and the prior Π is a probability measure on the measurable space $(\Theta, \mathscr{B})$. We assume that the distribution P_θ of X given θ is a *regular conditional distribution* on the sample space $(\mathfrak{X}, \mathscr{X})$ of the data, i.e. a *Markov kernel* from $(\Theta, \mathscr{B})$ into $(\mathfrak{X}, \mathscr{X})$:

(i) The map $A \mapsto P_\theta(A)$ is a probability measure for every $\theta \in \Theta$.
(ii) The map $\theta \mapsto P_\theta(A)$ is measurable for every $A \in \mathscr{X}$.

Then the pair (X, θ) has a well defined joint distribution on the product space $(\mathfrak{X} \times \Theta, \mathscr{X} \times \mathscr{B})$, given by

$$\Pr\bigl(X \in A, \theta \in B\bigr) = \int_B P_\theta(A)\, d\Pi(\theta).$$

This gives rise to the *marginal distribution* of X, defined by

$$\Pr(X \in A) = \int P_\theta(A)\, d\Pi(\theta).$$

By Kolmogorov's definition the conditional probabilities $\Pr(\theta \in B|\, X)$, for $B \in \mathscr{B}$, are always well defined, as measurable functions of X such that

$$\mathrm{E}\Pr(\theta \in B|\, X) 1_A(X) = \Pr(X \in A, \theta \in B), \qquad \text{every } A \in \mathscr{X}.$$

If the measurable space $(\Theta, \mathscr{B})$ is not too big, then there also exists a *regular version* of the conditional distribution: a Markov kernel from $(\mathfrak{X}, \mathscr{X})$ into $(\Theta, \mathscr{B})$. (See Section 2.1.1 below.) We shall consider the existence of a regular version necessary to speak of a true posterior distribution. A sufficient condition is that Θ is a Polish topological space [1] and $\mathscr{B}$ its Borel σ-field. More generally, it suffices that $(\Theta, \mathscr{B})$ is a *standard Borel space*, which by

[1] A topological space is called Polish if its topology is generated by a metric that makes it complete and separable.

definition is a measurable space that is isomorphic to a Polish space with its Borel σ-field. This condition will be met in all examples.

Even though the posterior distribution can thus usually be defined, some further care may be needed. It is inherent in the definition that the conditional probabilities $\Pr(\theta \in B|\, X)$ are unique only up to null sets under the marginal distribution of X. Using a regular version (on a standard Borel space) limits these null sets to a single null set that works for every measurable set B, but does not eliminate them altogether. This is hardly a concern if the full Bayesian setup is adopted, as this defines the marginal distribution of X as the appropriate data distribution. However, if the Bayesian framework is viewed as a method for inference only, and it is allowed that the "true" data X is generated according to some distribution P_0 different from the marginal distribution of X in the Bayesian setup, then the exceptional "null sets" may well have nonzero mass under this "true" distribution, and it is impossible to speak of *the* posterior distribution.

Obviously, this can only happen under serious "misspecification" of the prior. In particular, no problem arises if

$$P_0 \ll \int P_\theta \, d\Pi(\theta),$$

which is guaranteed for instance if P_0 is dominated by P_θ for θ in a set of positive prior probability. In parametric situations the latter condition is very reasonable, but the nonparametric case can be more subtle, particularly if the set of all P_θ is not dominated. Then there may be a "natural" way of defining the posterior distribution consistently for all X, but it must be kept in mind that this is not dictated by Bayes's rule alone. An important example of this situation arises with the nonparametric Dirichlet prior, where the marginal distribution is the normalized base measure, which may or may not dominate the distribution of the data.

For a dominated collection of measures P_θ it is generally possible to select densities p_θ relative to some σ-finite dominating measure μ such that the maps $(x, \theta) \mapsto p_\theta(x)$ are jointly measurable. Then a version of the posterior distribution is given by *Bayes's formula*

$$B \mapsto \frac{\int_B p_\theta(X) \, d\Pi(\theta)}{\int p_\theta(X) \, d\Pi(\theta)}. \tag{2.1}$$

Of course, this expression is defined only if the denominator $\int p_\theta(X)\, d\Pi(\theta)$, which is the *marginal density* of X, is positive. Definitional problems arise (only) if this is *not* the case under the true distribution of the data. Incidently, the formula also shows that a Polishness assumption on $(\Theta, \mathscr{B})$ is sufficient, but not necessary, for existence of the posterior distribution: (2.1) defines a Markov kernel as soon as it is well defined.

A related issue concerns the supports of prior and posterior. In a vague sense the support of a measure is a smallest set that contains all its mass. A precise definition is possible only under assumptions on the measurable space. We limit ourselves to Polish spaces, for which the following definition of support can be shown to be well posed.

Definition 2.1 [support] The *support* of a probability measure on the Borel sets of a Polish space is the smallest closed set of probability one. Equivalently, it is the set of all elements of the space for which every open neighbourhood has positive probability.

It is clear that a posterior distribution will not recover a "nonparametric" set of true distributions unless the prior has a large support. Later this will be made precise in terms of posterior consistency (at a rate), which of course depends both on the prior and on the way the data distribution P_θ depends on the parameter θ.

2.0.4 Absolute continuity

Bayes's formula (2.1) is available if the model $(P_\theta: \theta \in \Theta)$ is dominated. This is common in parametric modelling, but may fail naturally in nonparametric situations. As a consequence, sometimes we perform Bayesian analysis without Bayes. Mathematically this is connected to absolute continuity of prior and posterior distributions.

It seems natural that a prior distribution supported on a certain set yields a posterior distribution supported inside the same set. Indeed, the equality $\Pi(B) = \mathrm{E}\Pr(\theta \in B|\, X)$ immediately gives the implication: if $\Pi(B) = 1$, then $\Pr(\theta \in B|\, X) = 1$ almost surely. The exceptional null set is again relative to the marginal distribution of X, and it may depend on the set B. The latter dependence can be quite serious. In particular, the valid complementary implication: if $\Pi(B) = 0$, then $\Pr(\theta \in B|\, X) = 0$ almost surely, should not be taken as proof that the posterior is always absolutely continuous with respect to the prior. The nonparametric Dirichlet prior exemplifies this again, as the posterior is typically orthogonal to the prior.

Again these issues do not arise in the case that the collection of distributions P_θ is dominated. Formula (2.1) immediately shows that the posterior is absolutely continuous relative to the prior in this case (where it is assumed that the formula is well posed). This can also be reversed. In the following lemma we assume that the posterior distribution is taken a regular version, whence it is unique up to a null set.

Lemma 2.2 *If both $(\mathfrak{X}, \mathscr{X})$ and $(\Theta, \mathscr{B})$ are standard Borel spaces, then the set of posterior distributions $\Pr(\theta \in B|\, X = x)$, where $x \in \mathfrak{X}_0$ for a measurable set $\mathfrak{X}_0 \subset \mathfrak{X}$ of marginal probability 1, is dominated by a σ-finite measure if and only if the collection $\{P_\theta: \theta \in \Theta_0\}$ is dominated by a σ-finite measure, for some measurable set $\Theta_0 \subset \Theta$ with $\Pi(\Theta_0) = 1$. In this case the posterior distributions are dominated by the prior.*

Proof A collection of probability measures $\{P_\theta: \theta \in \Theta\}$ on a standard Borel space is dominated iff it is separable relative to the total variation distance, and in this case the measures permit densities $x \mapsto p_\theta(x)$ that are jointly measurable in (x, θ) (e.g. Strasser, 1985, Lemmas 4.6 and 4.1). Formula (2.1) then gives a version of the posterior distribution, which is dominated by the prior. Any other version differs from this version by at most a null set $\mathfrak{X}_0^c$.

The converse follows by interchanging the roles of x and θ. If the set of posterior distributions is dominated by a σ-finite measure, then they can be represented by conditional densities $\pi(\theta|\, x)$ relative to the dominating measure, measurable in (x, θ), and we can reconstruct a regular version of the conditional distribution of θ given x by (2.1), with the roles of θ and x interchanged, which is dominated. By assumption the original distributions P_θ give another regular version of this conditional distribution, and hence agree with the dominated version on a set of probability one. □

2.1 COMPLEMENTS

2.1.1 Regular versions

Given a random vector (θ, X) in a product space $(\Theta \times \mathfrak{X}, \mathscr{B} \times \mathscr{X})$ and $B \in \mathscr{B}$, the *conditional probability* $\Pr(\theta \in B |\, X)$ is defined as a measurable function $g(X)$ of X such that $\mathrm{E} g(X) 1_A(X) = \mathrm{E} 1_B(\theta) 1_A(X)$, for every $A \in \mathscr{X}$.

The existence of such a function can be proved using the Radon-Nikodym theorem. One notes that $A \mapsto \mathrm{E} 1_B(\theta) 1_A(X)$ defines a measure on $(\mathfrak{X}, \mathscr{X})$ that is absolutely continuous relative to the marginal distribution of X (if $\Pr(X \in A) = 0$, then $\mathrm{E} 1_B(\theta) 1_A(X) = 0$). Thus there exists a measurable function $g: \mathfrak{X} \to \mathbb{R}$ that gives the density of this measure relative to the marginal distribution of X.

The conditional probability is unique only up to a null set for the marginal distribution of X, as $\mathrm{E} g(X) 1_A(X) = \mathrm{E} \tilde{g}(X) 1_A(X)$, for every $\tilde{g}$ such that $g(X) = \tilde{g}(X)$, almost surely. If we define $\Pr(\theta \in B |\, X)$ for every set B, we thus leave indetermination for many null sets, whose union may well not be null. We call *regular version* of the conditional probability any map $(x, B) \mapsto G(x, B)$ such that

(i) $G(X, B)$ is a version of $\Pr(\theta \in B |\, X)$ for every B.
(ii) $B \mapsto G(x, B)$ is a probability measure, for every $x \in \mathfrak{X}$.

These conditions imply that $(x, B) \mapsto G(x, B)$ is a Markov kernel.

Sufficient for existence of such a regular version is that there are “not too many sets” B. Topological conditions can make this precise. See a book on measure-theoretic probability for the following two results.

Proposition 2.3 *If $(\Theta, \mathscr{B})$ is a Polish space with its Borel σ-field, then there exists a regular version of* $\Pr(\theta \in B |\, X)$.

An equivalent formulation solely in terms of measurable spaces, is as follows.

Proposition 2.4 *If Q is a probability distribution on a product measurable space $(\Theta \times \mathfrak{X}, \mathscr{B} \times \mathscr{X})$, where $(\Theta, \mathscr{B})$ is a Polish space with its Borel σ-field, then there exists a Markov kernel $(x \times B) \mapsto Q_{1|2}(x, B)$ such that $Q(A \times B) = \int_A Q_{1|2}(x, B)\, dQ_2(x)$, for Q_2 the second marginal distribution of Q.*

Exercises

2.1 (Extended Bayes’s rule) Suppose the distribution of X given θ is the convex combination of two orthogonal distributions with weights w_1 and $w_2 = 1 - w_1$ not depending on θ and densities $p_1(\cdot; \theta)$ and $p_2(\cdot; \theta)$ relative to dominating measures μ_1 and μ_2, respectively. Show that the posterior distribution of θ given X satisfies $d\Pi(\theta |\, X) \propto \bigl(w_1 p_1(X; \theta) 1_{S_1}(X) + w_2 p_2(X; \theta) 1_{S_2}(X)\bigr)\, d\Pi(\theta)$, for S_1 and S_2 disjoint measurable sets such that $\int_{S_2} p_1(x; \theta)\, d\mu_1(x) = \int_{S_1} p_2(x; \theta)\, d\mu_2(x) = 0$.

2.2 (Support) Show that the support of a probability measure on the Borel sets of a Polish space is well defined: there exists a smallest closed set of mass 1.

3

Priors on spaces of probability measures

In the nonparametric setting placing a prior distribution directly on the law of the data is natural. However, this comes with some technical complications. To limit these as much as possible we assume that the sample space $(\mathfrak{X}, \mathscr{X})$ is a Polish space with its Borel σ-field (the smallest σ-field containing all open sets), and consider priors on the collection $\mathfrak{M} = \mathfrak{M}(\mathfrak{X})$ of all probability measures on $(\mathfrak{X}, \mathscr{X})$.

3.1 Random measures, measurability

A prior Π on $\mathfrak{M}$ is a probabiity measure on a σ-field of subsets of $\mathfrak{M}$. Alternatively, it can be viewed as the law of a *random measure* P (a map from some probability space into $\mathfrak{M}$), and can be identified with the collection of "random probabilities" $P(A)$ of sets $A \in \mathscr{X}$. It is natural to choose the measurability structure on $\mathfrak{M}$ so that at least each $P(A)$ is a random variable, i.e. so that $\bigl(P(A)\colon A \in \mathscr{X}\bigr)$ is a *stochastic process* on the underlying probability space. We define the σ-field $\mathscr{M}$ on $\mathfrak{M}$ as the minimal one to make this true: $\mathscr{M}$ is equal to the smallest σ-field that makes all maps $M \mapsto M(A)$ from $\mathfrak{M}$ to $\mathbb{R}$ measurable, for $A \in \mathscr{X}$, and consider priors Π that are measures on $(\mathfrak{M}, \mathscr{M})$. Although other measurability structures are possible, the σ-field $\mathscr{M}$ is attractive for two important properties.

First it is identical to the Borel σ-field for the weak topology on $\mathfrak{M}$ (the topology of convergence in distribution in this space, see Proposition 3.1 at the end of this section). As $\mathfrak{M}$ is Polish under the weak topology (see Proposition 3.2 at the end of this section), this means that $(\mathfrak{M}, \mathscr{M})$ is a standard Borel space. As is noted in Section 2, this is desirable for the definition of posterior distributions, and also permits to speak of the support of a prior, called the *weak support* in this situation. Furthermore, the parameter θ from Section 2 that indexes the statistical model $(P_\theta\colon \theta \in \Theta)$ can be taken equal to the distribution P itself, with $\mathfrak{M}$ (or a subset) as the parameter set, giving a model of the form $(P\colon P \in \mathfrak{M})$. With respect to the σ-field $\mathscr{M}$ on the parameter set $\mathfrak{M}$ the data distributions are trivially "regular conditional probabilities":

(i) $P \mapsto P(A)$ is $\mathscr{M}$-measurable for every $A \in \mathscr{X}$,
(ii) $A \mapsto P(A)$ is a probability measure for every $P \in \mathfrak{M}$.

This mathematically justifies speaking of "drawing a measure P from the prior Π and next sampling observations X from P".

Second, the fact that $\mathscr{M}$ is generated by all maps $M \mapsto M(A)$, for $A \in \mathscr{X}$, implies that a map $P\colon (\Omega, \mathscr{U}, \Pr) \to (\mathfrak{M}, \mathscr{M})$ defined on some probability space is measurable precisely if the induced measure $P(A)$ of every set $A \in \mathscr{X}$ is a random variable. (See

Exercise 3.1.) Thus, as far as measurability goes, a random probability measure can be identified with a random element $\bigl(P(A)\colon A \in \mathscr{X}\bigr)$ in the product space $\mathbb{R}^{\mathscr{X}}$ (or $[0,1]^{\mathscr{X}}$).

Proposition 3.1 *If $\mathfrak{X}$ Polish, then the Borel σ-field $\mathscr{M}$ for the weak topology is also:*

(i) *the smallest σ-field on $\mathfrak{M}$ making all maps $P \mapsto P(A)$ measurable, for $A \in \mathscr{X}$;*
(ii) *the smallest σ-field on $\mathfrak{M}$ making all maps $P \mapsto P(A)$ measurable, for A in a generator $\mathscr{X}_0$ for $\mathscr{X}$;*
(iii) *the smallest σ-field on $\mathfrak{M}$ making all maps $P \mapsto \int \psi\, dP$ measurable, for $\psi \in \mathfrak{C}_b(\mathfrak{X})$.*

Consequently, a finite measure on $(\mathfrak{M}, \mathscr{M})$ is completely determined by the set of distributions induced under the maps $P \mapsto \bigl(P(A_1), \ldots, P(A_k)\bigr)$, for $A_1, \ldots, A_k \in \mathscr{X}_0$ and $k \in \mathbb{N}$; and also under the maps $P \mapsto \int \psi\, dP$, for $\psi \in \mathfrak{C}_b(\mathfrak{X})$.

Proposition 3.2 *The weak topology $\mathscr{W}$ on the set $\mathfrak{M}(\mathfrak{X})$ of Borel measures on a Polish space $\mathfrak{X}$ is Polish.*

3.2 Discrete random measures

Given a random vector $(N, W_{1,N}, \ldots, W_{N,N}, \theta_{1,N}, \ldots, \theta_{N,N})$, where $N \in \mathbb{N} \cup \{\infty\}$, $W_{1,N}, \ldots, W_{N,N}$ are nonnegative random variables with $\sum_{i=1}^N W_{i,N} = 1$, and $\theta_{1,N}, \ldots, \theta_{N,N}$ are random variables taking their values in $(\mathfrak{X}, \mathscr{X})$, we can define a random probability measure by

$$P = \sum_{i=1}^{N} W_{i,N} \delta_{\theta_{i,N}}.$$

The realizations of this prior are discrete with finitely or countably many support points, which may be different for each realization. Given the number N of support points, their "weights" $W_{1,N}, \ldots, W_{N,N}$ and "locations" $\theta_{1,N}, \ldots, \theta_{N,N}$ are often chosen independent.

Lemma 3.3 *If the support of N is unbounded and given N the weights and locations are independent, with given $N = n$, the weights having full support $\mathbb{S}_n$ and the locations full support $\mathfrak{X}^n$, for every n, then P has full support $\mathfrak{M}$.*

Proof Because the finitely discrete distributions are weakly dense in $\mathfrak{M}$, it suffices to show that P gives positive probability to any weak neighbourhood of a distribution $P^* = \sum_{i=1}^k w_i^* \delta_{\theta_i^*}$ with finite support. All distributions $P' := \sum_{i=1}^k w_i \delta_{\theta_i}$ with $(w_1, \ldots, w_k)$ and $(\theta_1, \ldots, \theta_k)$ sufficiently close to $(w_1^*, \ldots, w_k^*)$ and $(\theta_1^*, \ldots, \theta_k^*)$ are in such a neighbourhood. So are the measures $P' = \sum_{i=1}^\infty w_i \delta_{\theta_i}$ with $\sum_{i>k} w_i$ sufficiently small and $(w_1, \ldots, w_k)$ and $(\theta_1, \ldots, \theta_k)$ sufficiently close to their targets, as before.

If N is not identically infinite, then the assertion follows from the assumed positive probability of the events $\bigl\{N = k', \max_{i \le k'} |W_{i,k'} - w_i^*| \vee |\theta_{i,k'} - \theta_i^*| < \epsilon\bigr\}$ for every $\epsilon > 0$ and some $k' > k$, where we define $w_i^* = 0$ and θ_i^* arbitrarily for $k < i \le k'$.

If N is infinite with probability one, then the assertion follows similarly upon considering the events $\bigl\{\sum_{i>k} W_{i,\infty} < \epsilon, \max_{i \le k} |W_{i,k'} - w_i^*| \vee |\theta_{i,k} - \theta_i^*| < \epsilon\bigr\}$. These events have positive probability, as they refer to an open subset of $\mathbb{S}_\infty$. □

The prior is computationally more tractable if N is finite and bounded, but its full support is then not guaranteed. To achieve reasonable large sample properties, N must either depend on the sample size n, or be given a prior with infinite support.

An important special case is obtained by choosing $N \equiv \infty$, yielding a prior of the form

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i}.$$

Further specializations are to choose $\theta_1, \theta_2, \ldots$ an i.i.d. sequence in $\mathfrak{X}$, and to choose the weights $W_1, W_2, \ldots$ by the stick-breaking algorithm of Section 3.2.1. If the common distribution of the θ_i has support the full space $\mathfrak{X}$, and the stick-breaking weights are as in Lemma 3.4, then this prior has full support.

3.2.1 Stick breaking

"Stick-breaking" is a technique to construct a prior on infinite probability vectors $(p_1, p_2, \ldots)$. The problem in hand is to distribute the total mass 1, which we identify with a stick of length 1, randomly to each element of $\mathbb{N}$. We first break the stick at a point given by the realization of a random variable $0 \leq Y_1 \leq 1$ and assign mass Y_1 to $1 \in \mathbb{N}$. We think of the remaining mass $1 - Y_1$ as a new stick, and break it into two pieces of relative lengths Y_2 and $1 - Y_2$ according to the realized value of another random variable $0 \leq Y_2 \leq 1$. We assign mass $(1-Y_1)Y_2$ to the point 2, and are left with a new stick of length $(1-Y_1)(1-Y_2)$. Continuing in this way, we assign mass to the point j equal to

$$p_j = \Bigl(\prod_{l=1}^{j-1} (1 - Y_l)\Bigr) Y_j. \tag{3.1}$$

Clearly, by continuing to infinity, this scheme will attach a random subprobability distribution to $\mathbb{N}$ for any sequence of random variables $Y_1, Y_2, \ldots$ with values in $[0, 1]$. Under mild conditions the probabilities p_j will sum to one.

Lemma 3.4 *The random subprobability distribution $(p_1, p_2, \ldots)$ lies in $\mathbb{S}_\infty$ almost surely iff $\mathrm{E}\bigl[\prod_{l=1}^{j}(1 - Y_l)\bigr] \to 0$ as $j \to \infty$. For independent variables $Y_1, Y_2, \ldots$ this condition is equivalent to $\sum_{l=1}^{\infty} \log \mathrm{E}(1 - Y_l) = -\infty$. In particular, for i.i.d. variables $Y_1, Y_2, \ldots$ it suffices that $\mathrm{P}(Y_1 > 0) > 0$. If for every $k \in \mathbb{N}$ the support of $(Y_1, \ldots, Y_k)$ is $(0, \infty)^k$, then the support of $(p_1, p_2, \ldots)$ is the whole space $\mathbb{S}_\infty$.*

Proof By induction, it easily follows that the leftover mass at stage j is equal to $1 - \sum_{l=1}^{j} p_l = \prod_{l=1}^{j}(1 - Y_l)$. Hence the random subprobability distribution will lie in $\mathbb{S}_\infty$ a.s. iff $\prod_{l=1}^{j}(1 - Y_l) \to 0$ a.s.. Since the leftover sequence is decreasing, nonnegative and bounded by 1, the almost sure convergence is equivalent to convergence in mean. If the Y_j's are independent, then this condition becomes $\prod_{l=1}^{j}(1 - \mathrm{E}(Y_l)) \to 0$ as $j \to \infty$, which is equivalent to the condition $\sum_{l=1}^{\infty} \log \mathrm{E}(1 - Y_l) = -\infty$.

The last assertion follows, because the probability vector $(p_1, \ldots, p_k)$ is a continuous function of $(Y_1, \ldots, Y_k)$, for every k. □

3.3 Random measures as stochastic processes

A general method of construction a random measure is to start with the stochastic process $\big(P(A)\colon A \in \mathscr{X}\big)$, constructed using Kolmogorov's consistency theorem (see Proposition 3.7), and next show that this process can be realized within $\mathfrak{M}$, viewed as a subset of $\mathbb{R}^{\mathscr{X}}$. As the properties of measures are much richer than can be described by the finite-dimensional distributions involved in Kolmogorov's theorem, this approach is non-trivial, but it can be pushed through by standard arguments. The detais are as follows.

For every finite collection $A_1, \ldots, A_k$ of Borel sets in $\mathfrak{X}$ the vector of probabilities $\big(P(A_1), \ldots, P(A_k)\big)$ obtained from a random measure P is an ordinary random vector in $\mathbb{R}^k$. The construction of P may start with the specification of the distributions of all vectors of this type. A simple, important example would be to specify these as Dirichlet distributions with parameter vector $\big(\alpha(A_1), \ldots, \alpha(A_k)\big)$, for a given Borel measure α. For any *consistent* specification of the distributions, Kolmogorov's theorem allows to construct on a suitable probability space (Ω, U, Pr) a stochastic process $\big(P(A)\colon A \in \mathscr{X}\big)$ with the given finite-dimensional distributions. If the marginal distributions correspond to those of a random measure, then it will be true that

(i) $P(\emptyset) = 0$, $P(\mathscr{X}) = 1$, a.s.
(ii) $P(A_1 \cup A_1) = P(A_1) + P(A_2)$, a.s., for any disjoint A_1, A_2.

Assertion (i) follows, because the distributions of $P(\emptyset)$ and $P(\mathscr{X})$ will be specified to be degenerate at 0 and 1, respectively, while (ii) can be read off from the degeneracy of the joint distribution of the three variables $P(A_1)$, $P(A_2)$ and $P(A_1 \cup A_2)$. Thus the process $\big(P(A)\colon A \in \mathscr{X}\big)$ will automatically define a *finitely-additive* measure on $(\mathfrak{X}, \mathscr{X})$.

A problem is that the exceptional null sets in (ii) might depend on the pair (A_1, A_2). If restricted to a countable subcollection $\mathscr{X}_0 \subset \mathscr{X}$ there would only be countably many pairs and the null sets could be gathered in a single null set. Then still when extending (ii) to σ-additivity, which is typically possible by similar distributional arguments, there would be uncountably many sequences of sets. This problem can be overcome through existence of a *mean measure*

$$\mu(A) = \mathrm{E}P(A).$$

For a valid random measure P, this necessarily defines a Borel measure on $\mathfrak{X}$. Existence of a mean measure is also sufficient for existence of version of $\big(P(A)\colon A \in \mathscr{X}\big)$ that is a measure on $(\mathfrak{X}, \mathscr{X})$.

Theorem 3.5 *Suppose that $\big(P(A)\colon A \in \mathscr{X}\big)$ is a stochastic process that satisfies (i) and (ii) and whose mean $A \mapsto \mathrm{E}P(A)$ is a Borel measure on $\mathfrak{X}$. Then there exists a version of P that is a random measure on $(\mathfrak{X}, \mathscr{X})$. More precisely, there exists a measurable map $\tilde{P}\colon (\Omega, U, \mathrm{Pr}) \to (\mathfrak{M}, \mathscr{M})$ such that $P(A) = \tilde{P}(A)$ almost surely, for every $A \in \mathscr{X}$.*

Proof Let $\mathscr{X}_0$ be a countable field that generates the Borel σ-field $\mathscr{X}$, denumerated arbitrarily as $A_1, A_2, \ldots$. Because the mean measure $\mu(A)\colon= \mathrm{E}P(A)$ is regular, there exists for every $i, m \in \mathbb{N}$ a compact set $K_{i,m} \subset A_i$ with $\mu(A_i - K_{i,m}) < 2^{-2i-2m}$. By Markov's inequality

$$\mathrm{Pr}\big(P(A_i - K_{i,m}) > 2^{-i-m}\big) \leq 2^{i+m}\mathrm{E}P(A_i - K_{i,m}) \leq 2^{-i-m}.$$

Consequently, the event $\Omega_m = \cap_i\{P(A_i - K_{i,m}) \leq 2^{-i-m}\}$ possesses probability at least $1 - 2^{-m}$, and $\liminf \Omega_m$ possesses probability 1, by the Borel-Cantelli lemma.

Because $\mathscr{X}_0$ is countable, the null sets involved in (i)-(ii) with $A_1, A_2 \in \mathscr{X}_0$ can be aggregated into a single null set N. For every $\omega \notin N$ the process P is a finitely additive measure on $\mathscr{X}_0$, with the resulting usual properties of monotonicity and sub-additivity. By increasing N if necessary we can also ensure that it is sub-additive on all finite unions of sets $A_i - K_{i,m}$.

Let $A_{i_1} \supset A_{i_2} \supset \cdots$ be an arbitrary decreasing sequence of sets in $\mathscr{X}_0$ with empty intersection. Then, for every fixed m, the corresponding compacts $K_{i_j,m}$ possess empty intersection also, whence there exists a finite J_m such that $\cap_{j\leq J_m} K_{i_j,m} = \emptyset$. This implies that

$$A_{i_{J_m}} = \cap_{j=1}^{J_m} A_{i_j} - \cap_{j=1}^{J_m} K_{i_j,m} \subset \cup_{j=1}^{J_m}(A_{i_j} - K_{i_j,m}).$$

Consequently, on the event $\Omega_m - N$,

$$\limsup_j P(A_{i_j}) \leq P(A_{i_{J_m}}) \leq \sum_{j=1}^{J_m} P(A_{i_j} - K_{i_j,m}) \leq 2^{-m}.$$

Thus on the event $\Omega_0 = \liminf \Omega_m - N$ the limit is zero. We conclude that for every $\omega \in \Omega_0$, the restriction of $A \mapsto P(A)$ to $\mathscr{X}_0$ is countably additive. By Carathéodory's theorem it extends to a measure $\tilde{P}$ on $\mathscr{X}$.

By construction $\tilde{P}(A) = P(A)$ almost surely, for every $A \in \mathscr{X}_0$. In particular $\mathrm{E}\tilde{P}(A) = \mathrm{E}P(A) = \mu(A)$, for every A in the field $\mathscr{X}_0$, whence by uniqueness of extension the mean measure of $\tilde{P}$ coincides with the original mean measure μ on $\mathscr{X}$. For every $A \in \mathscr{X}$, there exists a sequence $\{A_m\} \subset \mathscr{X}_0$ such that $\mu(A \,\Delta\, A_m) \to 0$. Then both $P(A_m \,\Delta\, A)$ and $\tilde{P}(A_m \,\Delta\, A)$ tend to zero in mean. Finite-additivity of P gives that $\bigl|P(A_m) - P(A)\bigr| \leq P(A_m \,\Delta\, A)$, almost surely, and by σ-additivity the same is true for $\tilde{P}$. This shows that $\tilde{P}(A) = P(A)$ almost surely, for every $A \in \mathscr{X}$.

This also proves that $\tilde{P}(A)$ is a random variable for every $A \in \mathscr{X}$, whence $\tilde{P}$ is a measurable map in $(\mathfrak{M}, \mathscr{M})$. □

Rather than from the process $\bigl(P(A)\colon A \in \mathscr{X}\bigr)$ indexed by all Borel sets, we may wish to start from a smaller set $\bigl(P(A)\colon A \in \mathscr{X}_0\bigr)$ of variables, for some $\mathscr{X}_0 \subset \mathscr{X}$. As shown in the proof of the preceding theorem a countable collection $\mathscr{X}_0$ suffices, but compact sets play a special role.

Theorem 3.6 *Suppose that $\bigl(P(A)\colon A \in \mathscr{X}_0\bigr)$ is a stochastic process that satisfies (i) and (ii) for a countable field $\mathscr{X}_0$ that generates $\mathscr{X}$ and is such that for every $A \in \mathscr{X}_0$ and $\epsilon > 0$ there exists a compact $K_\epsilon \subset \mathfrak{X}$ and $A_\epsilon \in \mathscr{X}_0$ such that $A_\epsilon \subset K_\epsilon \subset A$ and $\mu(A - A_\epsilon) < \epsilon$, where μ is the mean $\mu(A) = \mathrm{E}P(A)$. Then there exists a random measure that extends P to $\mathscr{X}$.*

The proof of the theorem follows the same lines, except that, if the compacts K_ϵ are not elements of $\mathscr{X}_0$, the bigger sets $A - A_\epsilon$ must be substituted for $A - K_\epsilon$ when bounding the P-measure of this set.

For instance, for $\mathfrak{X} = \mathbb{R}^k$ we can choose $\mathscr{X}_0$ equal to the finite unions of cells $(a, b]$, with the compacts and A_ϵ equal to the corresponding finite unions of the intervals $[a_\epsilon, b]$

and $(a_\epsilon, b]$ for a_ϵ descending to a. By restricting to rational endpoints we obtain a countable collection.

3.4 COMPLEMENTS

Proposition 3.7 *[Kolmogorov extension theorem] For every finite subset S of an arbitrary set T let P_S be a probability distribution on $\mathbb{R}^S$. Then there exists a probability space $(\Omega, \mathscr{U}, \mathrm{Pr})$ and measurable maps $X_t\colon \Omega \to \mathbb{R}$ such that $(X_t\colon t \in S) \sim P_S$ for every finite set S if and only if for every pair $S' \subset S$ of finite subsets $P_{S'}$ is the marginal distribution of P_S on $\mathbb{R}^{S'}$.*

For a proof see a book on measure-theoretic probability or stochastic processes.

Exercises

3.1 For each A in an arbitrary index set $\mathscr{A}$ let $f_A\colon \mathfrak{M} \to \mathbb{R}$ be an arbitrary map.

(a) Show that there exists a smallest σ-field $\mathscr{M}$ such that every map f_A is measurable.

(b) Show that a map $T\colon (\Omega, \mathscr{U}) \to (\mathfrak{M}, \mathscr{M})$ is measurable if and only if $f_A \circ T\colon \Omega \to \mathbb{R}$ is measurable for every $A \in \mathscr{A}$.

3.2 Let $K_1, K_2, \ldots$ be compact sets in a topological space such that $\cap_i K_i = \emptyset$. Show that $\cap_{i=1}^m K_i = \emptyset$ for some m.

3.3 Show that the discrete probability measures with finitely many support points are dense in the set of all Borel probability measures on a Polish space (or $\mathbb{R}^k$) relative to the weak topology.

3.4 Show that for any Borel set $A \subset \mathbb{R}^k$ and finite measure μ on the Borel sets, and every $\epsilon > 0$, there exists a compact set K with $K \subset A$ and $\mu(A - K) < \epsilon$. [Let $\mathscr{X}_0$ be the set of all Borel sets A such that there exists for every $\epsilon > 0$ a closed set F and open set G with $F \subset A \subset G$ and $\mu(G - F) < \epsilon$. Show that $\mathfrak{X}_o$ is a σ-field. Show that it is the Borel σ-field. Show that the sets F can be taken compact without loss of generality.]

3.5 Consider a stick-breaking scheme with independent variables Y_k with $1 - \mathrm{Pr}(Y_k = 0) = 1/k^2 = \mathrm{Pr}(Y_k = 1 - e^{-k})$. Show that stick is “not finished”: $\sum_k p_k < 1$ almost surely.

4

Dirichlet process

The Dirichlet process is the “normal distribution of Bayesian nonparametrics”. It is the default prior on spaces of probability measures, and a building block for priors on other structures.

4.1 Finite-dimensonal Dirichlet distribution

A random vector $X = (X_1, \ldots, X_k)$ with values in the k-dimensional unit simplex $\mathbb{S}_k\colon= \bigl\{(s_1, \ldots, s_k)\colon s_j \geq 0, \sum_{j=1}^k s_j = 1\bigr\}$ is said to possess a *Dirichlet distribution* with parameters $k \in \mathbb{N}$ and $\alpha_1, \ldots, \alpha_k > 0$ if it has density proportional to $x_1^{\alpha_1-1} \cdots x_k^{\alpha_k-1}$ with respect to the Lebesgue measure on $\mathbb{S}_k$.

The unit simplex $\mathbb{S}_k$ is a subset of a $(k-1)$-dimensional affine space, and so “its Lebesgue measure” is to be understood to be $(k-1)$-dimensional Lebesgue measure appropriately mapped to $\mathbb{S}_k$. The norming constant of the Dirichlet density depends on the precise construction. Alternatively, the vector X may be described through the vector $(X_1, \ldots, X_{k-1})$ of its first $k-1$ coordinates, the last coordinate being fixed by the relationship $X_k = 1 - \sum_{i=1}^{k-1} X_i$. This vector has density proportional to $x_1^{\alpha_1-1} \cdots x_{k-1}^{\alpha_{k-1}-1}(1 - x_1 - \cdots - x_{k-1})^{\alpha_k-1}$ with respect to the usual $(k-1)$-dimensional Lebesgue measure restricted to the set $\mathbb{D}_k = \{(x_1, \ldots, x_{k-1})\colon \min_i x_i \geq 0, \sum_{i=1}^{k-1} x_i \leq 1\}$. The inverse of the normalizing constant is given by the *Dirichlet form*

$$\begin{aligned}\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-x_1-\cdots-x_{k-2}} & x_1^{\alpha_1-1} x_2^{\alpha_2-1} \cdots x_{k-1}^{\alpha_{k-1}-1} \\ & \times (1 - x_1 - \cdots - x_{k-1})^{\alpha_k-1}\, dx_{k-1} \cdots dx_2\, dx_1.\end{aligned} \tag{4.1}$$

The Dirichlet distribution takes its name from this integral, which was can be evaluated to $\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)/\Gamma(\alpha_1 + \cdots + \alpha_k)$ by successive integrations and scalings to beta integrals.

Definition 4.1 (Dirichlet distribution) The *Dirichlet distribution* $\mathrm{Dir}(k; \alpha)$ with parameters $k \in \mathbb{N} - \{1\}$ and $\alpha = (\alpha_1, \ldots, \alpha_k) > 0$ is the distribution of a vector $(X_1, \ldots, X_k)$ such that $\sum_{i=1}^k X_i = 1$ and such that $(X_1, \ldots, X_{k-1})$ has density

$$\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}\, x_1^{\alpha_1} x_2^{\alpha_2-1} \cdots x_{k-1}^{\alpha_{k-1}-1}(1 - x_1 - \cdots - x_{k-1})^{\alpha_k-1}, \quad x \in \mathbb{D}_k. \tag{4.2}$$

The Dirichlet distribution with parameters k and $\alpha \geq 0$, where $\alpha_i = 0$ for $i \in I \subsetneq \{1, \ldots, k\}$, is the distribution of the vector $(X_1, \ldots, X_k)$ such that $X_i = 0$ for $i \in I$

and such that $(X_i \colon i \notin I)$ possesses a lower-dimensional Dirichlet distribution, given by a density of the form (4.2).

For $k = 2$ the vector (X_1, X_2) is completely described by a single coordinate, where $X_1 \sim \mathrm{Be}(\alpha_1, \alpha_2)$ and $X_2 = 1 - X_1 \sim \mathrm{Be}(\alpha_2, \alpha_1)$. Thus the Dirichlet distribution is a multivariate generalization of the Beta distribution. The $\mathrm{Dir}(k; 1, \ldots, 1)$-distribution is the uniform distribution on $\mathbb{S}_k$.

Throughout the section we write $|\alpha|$ for $\sum_{i=1}^k \alpha_i$.

Proposition 4.2 (Gamma representation) *If* $Y_i \overset{ind}{\sim} \mathrm{Ga}(\alpha_i, 1)$*, then* $(Y_1/Y, \ldots, Y_k/Y) \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$*, and is independent of and* $Y\colon = \sum_{i=1}^k Y_i$*.*

Proof We may assume that all α_i are positive. The Jacobian of the inverse of the transformation $(y_1, \ldots, y_k) \mapsto (y_1/y, \ldots, y_{k-1}/y, y) =\colon (x_1, \ldots, x_{k-1}, y)$ is given by $y^{k-1}(1 - x_1 - \cdots - x_{k-1})$. The density of the $\mathrm{Ga}(\alpha_i, 1)$-distribution is proportional to $e^{-y_i} y_i^{\alpha_i - 1}$. Therefore the joint density of $(Y_1/Y, \ldots, Y_{k-1}/Y, Y)$ is, proportional to,

$$e^{-y} y^{|\alpha|-1} x_1^{\alpha_1 - 1} \cdots x_{k-1}^{\alpha_{k-1}-1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}.$$

This factorizes into a Dirichlet density of dimension $k - 1$ and the $\mathrm{Ga}(|\alpha|, 1)$-density of Y. □

Proposition 4.3 (Aggregation) *If* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$ *and* $Z_j = \sum_{i \in I_j} X_i$ *for a given partition* $I_1, \ldots, I_m$ *of* $\{1, \ldots, k\}$*, then*

(i) $(Z_1, \ldots, Z_m) \sim \mathrm{Dir}(m; \beta_1, \ldots, \beta_m)$*, where* $\beta_j = \sum_{i \in I_j} \alpha_i$*, for* $j = 1, \ldots, m$*.*
(ii) $(X_i/Z_j \colon i \in I_j) \overset{ind}{\sim} \mathrm{Dir}(\#I_j; \alpha_i, i \in I_j)$*, for* $j = 1, \ldots, m$*.*
(iii) $(Z_1, \ldots, Z_m)$ *and* $(X_i/Z_j \colon i \in I_j, j = 1, \ldots, m)$ *are independent.*

Conversely, if X *is a random vector such that (i)–(iii) hold, for a given partition* $I_1, \ldots, I_m$ *and* $Z_j = \sum_{i \in I_j} X_i$*, then* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$*.*

Proof In terms of the Gamma representation $X_i = Y_i/Y$ of Proposition 4.2 we have

$$Z_j = \frac{\sum_{i \in I_j} Y_i}{Y}, \qquad \text{and} \qquad \frac{X_i}{Z_j} = \frac{Y_i}{\sum_{i \in I_j} Y_i}.$$

Because $W_j\colon = \sum_{i \in I_j} Y_i \overset{\text{ind}}{\sim} \mathrm{Ga}(\beta_j, 1)$ for $j = 1, \ldots, m$, and $\sum_j W_j = Y$, the Dirichlet distributions in (i) and (ii) are immediate from Proposition 4.2. The independence in (ii) is immediate from the independence of the groups $(Y_i \colon i \in I_j)$, for $j = 1, \ldots, m$. By Proposition 4.2 W_j is independent of $(Y_i/W_j \colon iinI_j)$, for every j, whence by the independence of the groups the variables $W_j, (Y_i/W_j \colon i \in I_j)$, for $j = 1, \ldots, m$, are jointly independent. Then (iii) follows, because $(X_i/Z_j \colon i \in I_j, j = 1, \ldots, m)$ is a function of $(Y_i/W_j \colon i \in I_j, j = 1, \ldots, m)$ and $(Z_1, \ldots, Z_m)$ is a function of $(W_j \colon j = 1, \ldots, m)$.

The converse also follows from the Gamma representation. □

Proposition 4.4 (Moments) *If* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$*, then* $X_i \sim \mathrm{Be}(\alpha_i, |\alpha| - \alpha_i)$*. In*

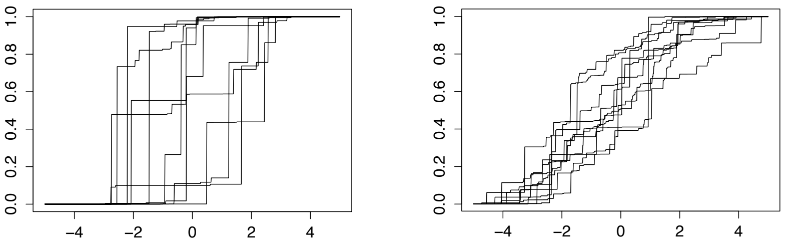

Figure 4.1 Cumulative distribution functions of 10 draws from the Dirichlet process with base measures $N(0,2)$ (left) and $10N(0,2)$ (right). (Computations based on Sethuraman representation truncated to 1000 terms.)

particular, $\mathrm{E}(X_i) = \alpha_i/|\alpha|$ *and* $\mathrm{var}(X_i) = \alpha_i(|\alpha| - \alpha_i)/(|\alpha|^2(|\alpha|+1))$. *Furthermore,* $\mathrm{cov}(X_i, X_j) = -\alpha_i\alpha_j/(|\alpha|^2(|\alpha|+1))$ *and, with* $r = r_1 + \cdots + r_k$,

$$\mathrm{E}(X_1^{r_1}\cdots X_k^{r_k}) = \frac{\Gamma(\alpha_1 + r_1)\cdots\Gamma(\alpha_k + r_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \times \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + r)}. \tag{4.3}$$

In particular, if $r_1, \ldots, r_k \in \mathbb{N}$, *then the expression in* (4.3) *is equal to* $\alpha_1^{[r_1]}\cdots\alpha_k^{[r_k]}/|\alpha|^{[r]}$, *where* $x^{[m]} = x(x+1)\cdots(x+m-1)$, $m \in \mathbb{N}$, *stands for the ascending factorial.*

Proof The first assertion follows from Proposition 4.3 by taking $m = 2$, $I_i = \{i\}$, $I_2 = I - \{i\}$, for $I = \{1, \ldots, k\}$. Next the expressions for expectation and variance follow by the properties of the beta distribution.

For the second assertion, we take $m = 2$, $I_1 = \{i, j\}$ and $I_2 = I - I_1$ in Proposition 4.3 to see that $X_i + X_j \sim \mathrm{Be}(\alpha_i + \alpha_j, |\alpha| - \alpha_i - \alpha_j)$. This gives $\mathrm{var}(X_i + X_j) = (\alpha_i + \alpha_j)(|\alpha| - \alpha_i - \alpha_j)/\big(|\alpha|^2(|\alpha|+1)\big)$, and allows to obtain the expression for the covariance from the identity $2\,\mathrm{cov}(X_i, X_j) = \mathrm{var}(X_i + X_j) - \mathrm{var}(X_i) - \mathrm{var}(X_j)$.

For the derivation of (4.3), observe that the mixed moment is the ratio of two Dirichlet forms (4.1) with parameters $(\alpha_1 + r_1, \ldots, \alpha_k + r_k)$ and $(\alpha_1, \ldots, \alpha_k)$. □

4.2 Dirichlet process

Definition 4.5 (Dirichlet process) A random measure P on $(\mathfrak{X}, \mathscr{X})$ is said to possess a *Dirichlet process* distribution $\mathrm{DP}(\alpha)$ with *base measure* α, if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$\big(P(A_1), \ldots, P(A_k)\big) \sim \mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big). \tag{4.4}$$

In this definition α is a given finite positive Borel measure on $(\mathfrak{X}, \mathscr{X})$. We write $|\alpha| = \alpha(\mathfrak{X})$ for its total mass and $\bar{\alpha} = \alpha/|\alpha|$ for the probability measure obtained by normalizing α, respectively, and use the notations $P \sim \mathrm{DP}(\alpha)$ and $P \sim \mathrm{DP}\big(|\alpha|, \bar{\alpha}\big)$ interchangeably to say that P has a Dirichlet process distribution with base measure α.

Existence of the Dirichlet process is not obvious, but proved below.

Definition 4.5 specifies the joint distribution of the vector $\big(P(A_1), \ldots, P(A_k)\big)$, for any measurable partition $\{A_1, \ldots, A_k\}$ of the sample space. In particular, it specifies the distribution of $P(A)$, for every measurable set A, and hence the *mean measure* $A \mapsto \mathrm{E}P(A)$. By Proposition 4.4,

$$\mathrm{E}(P(A)) = \bar{\alpha}(A).$$

Thus the mean measure is the normalized base measure $\bar{\alpha}$, which is a valid Borel measure by assumption. Therefore Theorem 3.5 implies existence of the Dirichlet process $\mathrm{DP}(\alpha)$ provided the specification of distributions can be consistently extended to any vector of the type $\big(P(A_1), \ldots, P(A_k)\big)$, for arbitrary measurable sets and not just partitions, in such a way that it gives a finitely-additive measure.

An arbitrary collection $A_1, \ldots, A_k$ of measurable sets defines a collection of 2^k atoms of the form $A_1^* \cap A_2^* \cap \cdots \cap A_k^*$, where A^* stands for A or A^c. These atoms $\{B_j : j = 1, \ldots, 2^k\}$ (some of which may be empty) form a partition of the sample space, and hence the joint distribution of $\big(P(B_j) : j = 1, \ldots, 2^k\big)$ is defined by Definition 4.5. Every A_i can be written as a union of atoms, and $P(A_i)$ can be defined accordingly as the sum of the corresponding $P(B_j)$'s. This defines the distribution of the vector $\big(P(A_1), \ldots, P(A_k)\big)$.

To prove the existence of a stochastic process $\big(P(A) : A \in \mathscr{X}\big)$ that possesses these marginal distributions, it suffices to verify that this collection of marginal distributions is consistent in the sense of Kolmogorov's extension theorem. Consider the distribution of the vector $\big(P(A_1), \ldots, P(A_{k-1})\big)$. This has been defined using the coarser partitioning in the 2^{k-1} sets of the form $A_1^* \cap A_2^* \cap \cdots \cap A_{k-1}^*$. Every set in this coarser partition is a union of two sets in the finer partition used previously to define the distribution of $\big(P(A_1), \ldots, P(A_k)\big)$. Therefore, consistency pertains if the distributions specified by Definition 4.5 for two partitions, where one is finer than the other, are consistent.

Let $\{A_1, \ldots, A_k\}$ be a measurable partition and let $\{A_{i1}, A_{i2}\}$ be a further measurable partition of A_i, for $i = 1, \ldots, k$. Then Definition 4.5 specifies that

$$\begin{aligned}&\big(P(A_{11}), P(A_{12}), P(A_{21}), \ldots, P(A_{k1}), P(A_{k2})\big) \\ &\qquad\qquad \sim \mathrm{Dir}\big(2k; \alpha(A_{11}), \alpha(A_{12}), \alpha(A_{21}), \ldots, \alpha(A_{k1}), \alpha(A_{k2})\big).\end{aligned}$$

In view of the group additivity of finite dimensional Dirichlet distributions given by Proposition 4.3, this implies

$$\Big(\sum_{j=1}^{2} P(A_{1j}), \ldots, \sum_{j=1}^{2} P(A_{kj})\Big) \sim \mathrm{Dir}\Big(k; \sum_{j=1}^{2} \alpha(A_{1j}), \ldots, \sum_{j=1}^{2} \alpha(A_{kj})\Big).$$

Consistency follows as the right side is $\mathrm{Dir}\big(k; \alpha(A_1), \ldots, \alpha(A_k)\big)$, since α is a measure.

That $P(\emptyset) = 0$ and $P(\mathfrak{X}) = 1$ almost surely follow from the fact that $\{\emptyset, \mathfrak{X}\}$ is an eligible partition in Definition 4.5, whence $\big(P(\emptyset), P(\mathfrak{X}),\big) \sim \mathrm{Dir}(2; 0, |\alpha|)$ by (4.4). That $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ almost surely for every disjoint pair of measurable sets A_1, A_2, follows similarly from consideration of the distributions of the vectors $\big(P(A_1), P(A_2), P(A_1^c \cap A_2^c)\big)$ and $\big(P(A_1 \cup A_2), P(A_1^c \cap A_2^c)\big)$, whose three and two components both add up to 1.

We have proved the existence of the Dirichlet process distribution $\mathrm{DP}(\alpha)$ for every Polish sample space and every base measure α.

4.3 The Sethuraman representation

The *Sethuraman representation* of the Dirichlet process is a random discrete measure of the type discussed in Section 3.2, with stick-breaking weights, as in Section 3.2.1, based on the Beta-distribution. The random support points are generated from the normalized base measure.

The representation gives an easy method to simulate a Dirichlet process, at least approximately. It also proves the remarkable fact that realizations from the Dirichlet measure are discrete measures, with probability one.

In view of the results of Section 3.2, we also infer that the Dirichlet process is fully supported relative to the weak topology.

Theorem 4.6 (Sethuraman) *If $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar\alpha$ and $Y_1, Y_2, \ldots \overset{iid}{\sim} \mathrm{Be}(1, M)$ are independent random variables and $V_j = Y_j \prod_{l=1}^{j-1}(1 - Y_l)$, then $\sum_{j=1}^\infty V_j \delta_{\theta_j} \sim \mathrm{DP}(M\bar\alpha)$.*

Proof Because $\mathrm{E}\bigl(\prod_{l=1}^{j}(1 - Y_l)\bigr) = (M/(M+1))^j \to 0$, the stick-breaking weights V_j form a probability vector a.s. (c.f. Lemma 3.4), so that P is a probability measure a.s..

For $j \geq 2$ define $V_j' = Y_j \prod_{l=2}^{j-1}(1 - Y_l)$ and $\theta_j' = \theta_{j+1}$. Then $V_j = (1 - Y_1)V_{j-1}'$ for every $j \geq 1$ and hence

$$P\colon = V_1\delta_{\theta_1} + \sum_{j=2}^\infty V_j \delta_{\theta_j} = Y_1\delta_{\theta_1} + (1 - Y_1)\sum_{j=1}^\infty V_j'\delta_{\theta_j'}.$$

The random measure $P'\colon = \sum_{j=1}^\infty V_j'\delta_{\theta_j'}$ has exactly the same structure as P, and hence possesses the same distribution. Furthermore, it is independent of (Y_1, θ_1).

We conclude that P satisfies the distributional equation (4.5) given below, and the theorem follows from Lemma 4.7. □

The distributional equation for the Dirichlet process used in the preceding proof is of independent interest. For independent random variables $Y \sim \mathrm{Be}\bigl(1, |\alpha|\bigr)$ and $\theta \sim \bar\alpha$, consider the equation

$$P =_d Y\delta_\theta + (1 - Y)P. \tag{4.5}$$

We say that a random measure P that is independent of (Y, θ) is a solution to equation (4.5) if for every measurable partition $\{A_1, \ldots, A_k\}$ of the sample space the random vectors obtained by evaluating the random measures on its left and right sides are equal in distribution in $\mathbb{R}^k$.

Lemma 4.7 *For given independent $\theta \sim \bar\alpha$ and $Y \sim \mathrm{Be}\bigl(1, |\alpha|\bigr)$, the Dirichlet process $\mathrm{DP}(\alpha)$ is the unique solution of the distributional equation (4.5).*

Proof For a given measurable partition $\{A_1, \ldots, A_k\}$, the equation requires that $Q\colon = \bigl(P(A_1), \ldots, P(A_k)\bigr)$ has the same distribution as the vector $YN + (1 - Y)Q$, for $N \sim \mathrm{MN}(1; \bar\alpha(A_1), \ldots, \bar\alpha(A_k))$ and (Y, N) independent of Q.

We first show that the solution is unique in distribution. Let (Y_n, N_n) be a sequence of i.i.d. copies of (Y, N), and for two solutions Q and Q' that are independent of this sequence and suitably defined on the same probability space, set $Q_0 = Q$, $Q_0' = Q'$, and recursively define $Q_n = Y_nN_n + (1 - Y_n)Q_{n-1}$, $Q_n' = Y_nN_n + (1 - Y_n)Q_{n-1}'$, for $n \in \mathbb{N}$. Then every

Q_n is distributed as Q and every Q'_n is distributed as Q', because each of them satisfies the distributional equation. Also

$$\|Q_n - Q'_n\| = |1 - Y_n|\,\|Q_{n-1} - Q'_{n-1}\| = \prod_{i=1}^{n} |1 - Y_i|\,\|Q - Q'\| \to 0$$

with probability 1, since the Y_i are i.i.d. and are in $(0, 1)$ with probability one. This forces the distributions of Q and Q' to agree.

To prove that the Dirichlet process is a solution let $W_0, W_1, \ldots, W_k \overset{\text{ind}}{\sim} \text{Ga}(\alpha_i, 1)$, $i = 0, 1, \ldots, k$, where $\alpha_0 = 1$. Then by Proposition 4.3 the vector (W_0, W), for $W = \sum_{i=1}^{k} W_i$, is independent of the vector $Q\colon = (W_1/W, \ldots, W_k/W) \sim \text{Dir}(k, \alpha_1, \ldots, \alpha_k)$. Furthermore, $Y\colon = W_0/(W_0+W) \sim \text{Be}(1, |\alpha|)$ and $(Y, (1-Y)Q) \sim \text{Dir}(k+1; 1, \alpha_1, \ldots, \alpha_k)$. Thus for any $i = 1, \ldots, k$, merging the 0th cell with the ith, we obtain from Proposition 4.3 that, with e_i the ith unit vector,

$$Ye_i + (1 - Y)Q \sim \text{Dir}(k; \alpha + e_i), \quad i = 1, \ldots, k. \tag{4.6}$$

This gives the conditional distribution of the vector $YN + (1 - Y)Q$ given $N = e_i$. It follows that $YN + (1 - Y)Q$ given N possesses a $\text{Dir}(k; \alpha + N)$-distribution, just as p given N in Proposition 4.8. Because also the marginal distributions of N in the two cases are the same, so must be the marginal distributions of $YN + (1 - Y)N$ and p, where the latter is $p \sim \text{Dir}(k; \alpha)$. □

Proposition 4.8 (Conjugacy) *If $p \sim \text{Dir}(k; \alpha)$ and $N|\,p \sim \text{MN}(n, k; p)$, then $p|\,N \sim \text{Dir}(k; \alpha + N)$.*

Proof If some coordinate α_i of α is zero, then the corresponding coordinate p_i of p is zero with probability one, and hence so is the coordinate N_i of N. After removing these coordinates we can work with densities. The product of the Dirichlet density and the multinomial likelihood is proportional to

$$p_1^{\alpha_1-1} \cdots p_k^{\alpha_k-1} \times p_1^{N_1} \cdots p_k^{N_k} = p_1^{\alpha_1+N_1-1} \cdots p_k^{\alpha_k+N_k-1}.$$

This is proportional to the density of the $\text{Dir}(k; \alpha_1 + N_1, \ldots, \alpha_k + N_k)$-distribution. □

4.3.1 Self-similarity

For a measure P and measurable set B, let $P_{|B}$ stand for the restriction measure $P_{|B}(A) = P(A \cap B)$, and P_B for the conditional measure $P_B(A) = P(A|\,B)$, for B with $P(B) > 0$.

Theorem 4.9 (Self-similarity) *If $P \sim \text{DP}(\alpha)$, then $P_B \sim \text{DP}(\alpha_{|B})$, and the variable and processes $P(B)$, $\big(P_B(A)\colon A \in \mathscr{X}\big)$ and $\big(P_{B^c}(A)\colon A \in \mathscr{X}\big)$ are mutually independent, for any $B \in \mathscr{X}$ such that $\alpha(B) > 0$.*

Proof Because $P(B) \sim \text{Be}\big(\alpha(B), \alpha(B^c)\big)$ the condition that $\alpha(B) > 0$ implies that $P(B) > 0$ a.s., so that the conditional probabilities given B are well defined.

For given partitions $A_1, \ldots, A_r$ of B and $C_1, \ldots, C_s$ of B^c, the vector

$$X\colon = \big(P(A_1), \ldots, P(A_r), P(C_1), \ldots, P(C_s)\big)$$

possesses a Dirichlet distribution $\mathrm{Dir}\bigl(r+s;\alpha(A_1),\ldots,\alpha(A_r),\alpha(C_1),\ldots,\alpha(C_s)\bigr)$. By Proposition 4.3 the four variables or vectors

$$Z_1 := \sum_{i=1}^{r} X_i, \quad Z_2 := \sum_{i=r+1}^{r+s} X_i, \quad (X_1/Z_1,\ldots,X_r/Z_1), \quad (X_{r+1}/Z_2,\ldots,Z_{r+s}/Z_2)$$

are mutually independent, and the latter two vectors have Dirichlet distributions with the restrictions of the original parameters. These are precisely the variables $P(B)$, $P(B^c)$, and vectors with coordinates $P_B(A_i)$ and $P_{B^c}(C_i)$. □

Theorem 4.9 shows that the Dirichlet process "localized" by conditioning to a set B is again a Dirichlet process, with base measure the restriction of the original base measure. Furthermore, processes at disjoint localities are independent of each other, and also independent of the "macro level" variable $P(B)$. Within any given locality, mass is further distributed according to a Dirichlet process, independent of what happens to the "outside world". This property may be expressed by saying that locally a Dirichlet process is like itself; in other words it is *self similar*.

Exercises

4.1 Show that if $P \sim \mathrm{DP}(\alpha)$ and $\psi\colon \mathfrak{X} \to \mathfrak{Y}$ is a measurable mapping, then $P \circ \psi^{-1} \sim \mathrm{DP}(\beta)$, for $\beta = \alpha \circ \psi^{-1}$.

4.2 Show that if $P \sim \mathrm{DP}(\alpha)$, then $\mathrm{E}\int \psi\, dP = \int \psi\, d\bar\alpha$, and $\mathrm{var} \int \psi\, dP = \int (\psi - \int \psi\, d\bar\alpha)^2\, d\bar\alpha/(1+|\alpha|)$, for any measurable function ψ for which the integrals make sense (e.g. bounded). [Hint: proof this first for $\psi = 1_A$.]

4.3 Let $0 = T_0 < T_1 < T_2 < \cdots$ be the events of a standard Poisson process and let $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} \bar\alpha$ and independent of $(T_1, T_2, \ldots)$. Show that

$$P = \sum_{k=1}^{\infty} (e^{-T_{k-1}} - e^{-T_k})\delta_{\theta_k}$$

follows a Dirichlet process $\mathrm{DP}(\bar\alpha)$. How can we change the prior precision to $M \neq 1$?

4.4 Let $F \sim \mathrm{DP}(MG)$ be a Dirichlet process on $\mathfrak{X} = \mathbb{R}$, for a constant $M > 0$ and probability distribution G, identified by its cumulative distribution function $x \mapsto G(x)$. So F can be viewed as a random cumulative distribution function. Define its median as any value m_F such that $F(m_F-) \leq 1/2 \leq F(m_F)$. Show that

$$\mathrm{Pr}\bigl(m_F \leq x\bigr) = \int_{1/2}^{1} \beta\bigl(u, MG(x), M(1-G(x)\bigr)\, du,$$

where $\beta(\cdot, \alpha, \beta)$ is the density of the Beta-distribution.

4.5 Simulate and plot the cumulative distribution function of the Dirichlet processes $F \sim \mathrm{DP}(\Phi)$, $F \sim \mathrm{DP}(0.1\Phi)$, and $F \sim \mathrm{DP}(10\Phi)$. Do the same with the Cauchy base measure. [Suggestion use Sethuraman's presentation. Cut the series at an appropriate point.]

5

Tail-free processes

Consider a sequence $\mathcal{T}_0 = \{\mathfrak{X}\}$, $\mathcal{T}_1 = \{A_0, A_1\}$, $\mathcal{T}_2 = \{A_{00}, A_{01}, A_{10}, A_{11}\}$, and so on, of measurable partitions of the sample space $\mathfrak{X}$, obtained by splitting every set in the preceding partition into two new sets. See Figure 5.

With $\mathcal{E} = \{0, 1\}$ and $\mathcal{E}^* = \cup_{m=0}^{\infty} \mathcal{E}^m$, the set of all finite strings $\varepsilon_1 \cdots \varepsilon_m$ of 0's and 1's, we can index the 2^m sets in the mth partition $\mathcal{T}_m$ by $\varepsilon \in \mathcal{E}^m$, in such a way that $A_\varepsilon = A_{\varepsilon 0} \cup A_{\varepsilon 1}$ for every $\varepsilon \in \mathcal{E}^*$. Here $\varepsilon 0$ and $\varepsilon 1$ are the extensions of the string ε with a single symbol 0 or 1; the empty string indexes $\mathcal{T}_0$. Let $|\varepsilon|$ stand for the length of a string ε, and let $\varepsilon\delta$ be the concatenation of two strings $\varepsilon, \delta \in \mathcal{E}^*$. The set of all finite unions of sets A_ε, for $\varepsilon \in \mathcal{E}^*$, forms a sub-field of the Borel sets. We assume throughout that the splits are chosen rich enough that this generates the Borel σ-field.

Because the probability of any A_ε must be distributed to its "offspring" $A_{\varepsilon 0}$ and $A_{\varepsilon 1}$, a probability measure P must satisfy the *tree additivity* requirement $P(A_\varepsilon) = P(A_{\varepsilon 0}) +$

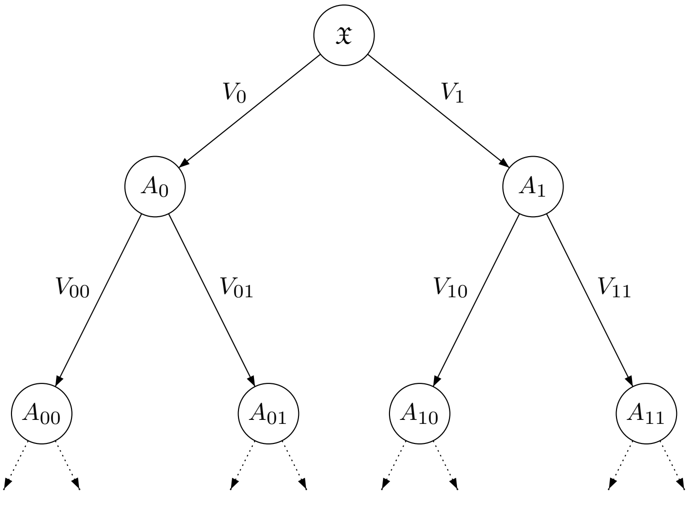

Figure 5.1 Tree diagram showing the distribution of mass over the first two partitions $\mathfrak{X} = A_0 \cup A_1 = (A_{00} \cup A_{01}) \cup (A_{10} \cup A_{11})$ of the sample space. Mass at a given node is distributed to its two childeren proportionally to the weights on the arrows. Every pair of V's on arrows originating from the same node add to 1.

$P(A_{\varepsilon 1})$. The relative weights of the offspring sets are the conditional probabilities

$$V_{\varepsilon 0} = P(A_{\varepsilon 0}|\, A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = P(A_{\varepsilon 1}|\, A_\varepsilon). \tag{5.1}$$

This suggests that, for given set $(V_\varepsilon\colon \varepsilon \in \mathcal{E}^*)$ of $[0,1]$-valued random variables, we might define a random measure P by

$$P(A_{\varepsilon_1\cdots\varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1\varepsilon_2} \cdots V_{\varepsilon_1\cdots\varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^m. \tag{5.2}$$

If $V_{\varepsilon 0} + V_{\varepsilon 1} = 1$ for every ε, then the stochastic process $\big(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^*\big)$ will satisfy the tree-additivity condition, and define a finitely additive measure on the field of all finite unions of sets A_ε, for $\varepsilon \in \mathcal{E}^*$.

Countable additivity is not immediate, but may be established using a mean measure, by the approach of Theorem 3.6.

Theorem 5.1 *Consider a sequence of partitions $\mathcal{T}_m = \{A_\varepsilon\colon \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets in $(\mathfrak{X}, \mathscr{X})$ and is such that every A_ε is the union of all $A_{\varepsilon\delta}$ whose closure is compact and satisfies $\overline{A}_{\varepsilon\delta} \subset A_\varepsilon$, where $\delta \in \mathcal{E}^*$. If $(V_\varepsilon\colon \varepsilon \in \mathcal{E}^*)$ is a stochastic process with $0 \le V_\varepsilon \le 1$ and $V_{\varepsilon 0} + V_{\varepsilon 1} = 1$ for all $\varepsilon \in \mathcal{E}^*$, and there exists a Borel measure μ such that $\mu(A_\varepsilon) = \mathrm{E} V_{\varepsilon_1} V_{\varepsilon_1\varepsilon_2} \cdots V_{\varepsilon_1\cdots\varepsilon_m}$, for every $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^*$, then there exists a random Borel measure P satisfying* (5.2).

Proof For fixed $\varepsilon \in \mathcal{E}^*$ there are at most countably many $A_{\varepsilon\delta}$ as stated, and their union is A_ε. Thus for any given $\eta > 0$ there exists a finite subcollection whose union $B_{\varepsilon,\eta}$ satisfies $\mu(A_\varepsilon - B_{\varepsilon,\eta}) < \eta$. The corresponding union $K_{\varepsilon,\eta}$ of the closures $\bar{A}_{\varepsilon\delta}$ is compact and satisfies $B_{\varepsilon,\eta} \subset K_{\varepsilon,\eta} \subset A_\varepsilon$. Thus we are in the situation of Theorem 3.6, with P defined by (5.2) as a finitely additive measure on the field consisting of all finite unions of A_ε. □

5.1 Tail-free processes

Consider a given partitioning tree $\mathcal{T}_1, \mathcal{T}_2, \ldots$ and a random measure P on the Borel sets, and define the splitting variables $(V_\varepsilon, \varepsilon \in \mathcal{E}^*)$ as in (5.1). Write $U \perp V$ to denote that random variables U and V are independent, and $U \perp V|\, Z$ to say that U and V are conditionally independent given a random variable Z.

Definition 5.2 (Tail-free) The random measure P is a *tail-free process* with respect to the sequence of partitions $\mathcal{T}_m$ if $\{V_0\} \perp \{V_{00}, V_{10}\} \perp \cdots \perp \{V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m\} \perp \cdots$.

A degenerate prior is certainly tail-free according to this definition (with respect to any sequence of partitions), since all its V-variables are degenerate at appropriate values. A nontrivial example is the Dirichlet process, as is seen in Theorem 5.3.

Theorem 5.3 *The* $\mathrm{DP}(\alpha)$ *prior is tail free. All splitting variables $V_{\varepsilon 0}$ are independent and $V_{\varepsilon 0} \sim \mathrm{Be}\big(\alpha(A_{\varepsilon 0}), \alpha(A_{\varepsilon 1})\big)$.*

Proof We must show that the vectors $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m)$ defined in (5.1) are mutually independent across levels m. It suffices to show sequentially for every m that this vector is independent of the vectors corresponding to lower levels. Because the vectors $(V_\varepsilon\colon \varepsilon \in \cup_{k\le m}\mathcal{E}^k)$, and $\big(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m\big)$ generate the same σ-field, it suffices to show that $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m)$ is independent of $\big(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m\big)$, for every fixed m.

This follows by an application of Proposition 4.3 to the vectors $\bigl(P(A_{\varepsilon\delta}): \varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E}\bigr)$ with the aggregation of the pairs $\bigl(P(A_{\varepsilon 0}), P(A_{\varepsilon 1})\bigr)$ into the sums $P(A_\varepsilon) = P(A_{\varepsilon 0}) + P(A_{\varepsilon 1})$.

The beta distributions also follow by Proposition 4.3 (and the fact that the first marginal of a $\mathrm{Dir}(2; \alpha, \beta)$ is a $\mathrm{Be}(\alpha, \beta)$-distribution). □

The mass $P(A_\varepsilon)$ of a partitioning set at level m can be expressed in the V-variables up to level m (see (5.2)), while, by their definition (5.1), the V-variables at higher levels control conditional probabilities. Therefore, tail-freeness makes the distribution of mass *within* every partitioning set in $\mathcal{T}_m$ independent of the distribution of the total mass one *among* the sets in $\mathcal{T}_m$. Definition 5.2 refers only to masses of partitioning sets, but under the assumption that the partitions generate the Borel sets, the independence extends to all Borel sets.

Lemma 5.4 *If P is a random measure that is tail-free relative to a sequence of partitions $\mathcal{T}_m = \{A_\varepsilon: \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets $\mathscr{X}$ in $\mathfrak{X}$, then for every $m \in \mathbb{N}$ the process $\bigl(P(A|\, A_\varepsilon): A \in \mathscr{X}, \varepsilon \in \mathcal{E}^m\bigr)$ is independent of the random vector $\bigl(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m\bigr)$.*

Proof Because P is a random measure, its mean measure $\mu(A) = \mathrm{E}P(A)$ is a well defined Borel probability measure. As $\mathcal{T} := \cup_m \mathcal{T}_m$ is a field, which generates the Borel σ-field by assumption, there exists for every $A \in \mathscr{X}$ a sequence A_n in $\mathcal{T}$ such that $\mu(A_n \,\Delta\, A) \to 0$. Because P is a random measure $P(A_n|\, A_\varepsilon) \to P(A|\, A_\varepsilon)$ in mean and hence a.s. along a subsequence. It follows that the random variable $P(A|\, A_\varepsilon)$ is measurable relative to the completion $\mathscr{U}_0$ of the σ-field generated by the variables $P(C|\, A_\varepsilon)$, for $C \in \mathcal{T}$. Every of latter variables is a finite sum of probabilities of the form $P(A_{\varepsilon\delta}|\, A_\varepsilon) = V_{\varepsilon\delta_1} \cdots V_{\varepsilon\delta_1\cdots\delta_k}$, for $\varepsilon \in \mathcal{E}^m$, $\delta = \delta_1 \cdots \delta_k \in \mathcal{E}^k$ and $k \in \mathbb{N}$. Therefore, by tail-freeness the σ-field $\mathscr{U}_0$ is independent of the σ-field generated by the variables $P(A_\varepsilon) = V_{\varepsilon_1} \cdots V_{\varepsilon_1\cdots\varepsilon_m}$, for $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^m$. □

Relative to the σ-field $\mathscr{M}$ generated by all maps $M \mapsto M(A)$ the process $\bigl(P(A|\, A_\varepsilon): A \in \mathcal{X}\bigr)$ contains all information about the conditional random measure $P(\cdot|\, A_\varepsilon)$. Thus the preceding lemma truly expresses that the "conditional measure within partitioning sets is independent of the distribution of mass among them".

Suppose that the data consists of an i.i.d. sample $X_1, \ldots, X_n$ from a distribution P, which is a-priori modelled as a tail-free process. For each $\varepsilon \in \mathcal{E}^*$, denote the number of observations falling in A_ε by

$$N_\varepsilon := \#\{1 \le i \le n: X_i \in A_\varepsilon\}. \tag{5.3}$$

For each m the vector $(N_\varepsilon: \varepsilon \in \mathcal{E}^m)$ collects the counts of all partitioning sets at level m. The following theorem shows that this vector contains all information (in the Bayesian sense) about the probabilities $\bigl(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m\bigr)$ of these sets: the additional information about the precise positions of the X_i within the partitioning sets is irrelevant.

Theorem 5.5 *If a random measure P is tail-free relative to a given sequence of partitions $\mathcal{T}_m = \{A_\varepsilon: \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets, then for every m and n the posterior distribution of $\bigl(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m\bigr)$ given an i.i.d. sample $X_1, \ldots, X_n$ from P is the same as the posterior distribution of this vector given $(N_\varepsilon: \varepsilon \in \mathcal{E}^m)$ defined in* (5.3)*, a.s..*

Proof We may generate the variables $P, X_1, \ldots, X_n$ in four steps:

(a) Generate the vector $\theta := \big(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m\big)$ from its prior.
(b) Given θ generate a multinomial vector $N = (N_\varepsilon: \varepsilon \in \mathcal{E}^m)$ with parameters n and θ.
(c) Generate the process $\eta := \big(P(A|\, A_\varepsilon): A \in \mathscr{X}, \varepsilon \in \mathcal{E}^m\big)$.
(d) Given (N, η) generate for every $\varepsilon \in \mathcal{E}^m$ a random sample of size N_ε from the measure $P(\cdot|\, A_\varepsilon)$, independently across $\varepsilon \in \mathcal{E}^m$, and $X_1, \ldots, X_n$ be the n values so obtained in a random order.

For a tail-free measure P step (c) is independent of step (a). Furthermore, the fact that step (b) uses only θ and not η means that $N \perp \eta|\, \theta$. Finally, that step (d) does not use θ can be expressed as $X \perp \theta|\, (N, \eta)$. Together these (in)dependencies imply that $\theta \perp X|\, N$, which is equivalent to the statement of theorem (see Exercise 5.3).

Thus the theorem is proved for this special representation of prior and data. Because the assertion depends on the joint distribution of (P, X, N) only, it is true in general. □

Given P the vector $(N_\varepsilon: \varepsilon \in \mathcal{E}^m)$ possesses a multinomial distribution with parameters n and $\big(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m\big)$. Finding the posterior distribution of the latter vector of cell probabilities therefore reduces to the finite dimensional problem of multinomial probabilities. This not only makes computations easy, but also means that asymptotic properties of the posterior distribution follow those of parametric problems, for instance easily leading to consistency in an appropriate sense. The result also justifies the term "tail-free" in that posterior computation can be carried out without looking at the tail of the prior.

Tail-free processes form a conjugate class of priors, in the sense that the posterior process is again tail-free.

Theorem 5.6 (Conjugacy) *The posterior process corresponding to observing an i.i.d. sample $X_1, \ldots, X_n$ from a distribution P that is a-priori modelled by a tail-free prior is tail-free (with respect to the same sequence of partitions as in the definition of the prior).*

Proof We must show that the vectors $(V_{\varepsilon 0}: \varepsilon \in \mathcal{E}^m)$ defined in (5.1) are mutually conditionally independent across levels m, given the data. As noted in the proof of Theorem 5.3, this is the case if $(V_{\varepsilon 0}: \varepsilon \in \mathcal{E}^m)$ is conditionally independent of $\big(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m\big)$, for every fixed m.

Together these vectors are equivalent to the vector $\big(P(A_\varepsilon): \varepsilon \in \mathcal{E}^{m+1}\big)$. Therefore, by Theorem 5.5 the joint posterior distribution of the latter vectors depends only on the cell counts, $N = (N_{\varepsilon\delta}: \varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E})$, and "conditionally given the data" can be interpreted as "given this vector N". Writing $V = (V_{\varepsilon\delta}: \varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E})$ and $\theta = (\theta_\varepsilon: \varepsilon \in \mathcal{E}^m)$, for $\theta_\varepsilon = P(A_\varepsilon)$, we can write the likelihood for (V, θ, N) as

$$\binom{n}{N} \prod_{\varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E}} (\theta_\varepsilon V_{\varepsilon\delta})^{N_{\varepsilon\delta}}\, d\Pi_1(V)\, d\Pi_2(\theta).$$

Here Π_1 and Π_2 are the marginal (prior) distributions of V and θ, and we have used that these vectors are independent under the assumption that P is tail-free. Clearly the likelihood factorizes in parts involving (V, N) and involving (θ, N). This shows that V and θ are conditionally independent given N. □

Exercises

5.1 Let G be a given continuous probability measure on $\mathbb{R}$, identified by its cdf. Let the partition at level m consist of the sets $\big(G^{-1}\big((i-1)2^{-m}\big), G^{-1}(i2^{-m})\big]$, for $i = 1, 2\dots, 2^m$. Let the variables $V_{\varepsilon 0}$ be independent with mean $1/2$ and define a random probability measure by (5.2). Find $\mathrm{E}P(A)$, for a given measurable set.

5.2 Suppose that $N \sim \mathrm{MN}(1, p_1, \dots, p_k)$ and given $N = e_j$ let X be drawn from a given probability measure P_j. Show that $X \sim \sum_j p_j P_j$. What is this latter measure if $p_j = P(A_j)$ and $P_j = P(\cdot|\, A_j)$ for a given measure P and measurable partition $\mathfrak{X} = \cup_j A_j$?

5.3 Let θ, η, N, X be random elements defined on a common probability space with values in Polish spaces. Show that

(a) $\eta \perp \theta$ if and only if $\Pr(\eta \in A|\, \theta) = \Pr(\eta \in A)$ almost surely, for all measurable sets A.

(b) $\theta \perp X|\, N$ if and only if $\Pr(\theta \in A|\, N) = \Pr(\theta \in A|\, X, N)$ almost surely for all measurable sets A.

(c) $\eta \perp (\theta, N)$ if and only if ($\eta \perp \theta|\, N$ and $\eta \perp N$).

(d) if $X \perp \theta|\, (N, \eta)$ and $\theta \perp \eta|\, N$, then $\theta \perp X|\, N$.

Conclude that if $\eta \perp (N, \theta)$ and $X \perp \theta|\, (N, \eta)$, then $\theta \perp X|\, N$. [The Polish assumptions guarantee that conditional distributions are well defined. Conditional independence of X and Y given Z means that $\Pr(X \in A, Y \in B|\, Z) = \Pr(X \in A|\, Z)\Pr(Y \in B|\, Z)$ almost surely, for every measurable sets A, B. The conditional expectation $\Pr(X \in A|\, Z)$ is a measurable function of Z such that $\mathrm{E}\Pr(X \in A|\, Z)1_C(Z) = \Pr(X \in A, Z \in C)$ for every measurable sets A, C.]

6

Dirichlet process (2)

Consider observations $X_1, X_2, \dots, X_n$ sampled independently from a distribution P that was drawn from a Dirichlet prior distribution, i.e.

$$P \sim \mathrm{DP}(\alpha), \qquad\qquad X_1, X_2, \dots \,|\, P \overset{\text{iid}}{\sim} P.$$

By an abuse of language, which we shall follow, such observations are often termed a *sample from the Dirichlet process*.

6.1 Posterior distribution

One of the most remarkable properties of the Dirichlet process prior is that the posterior distribution is again Dirichlet.

Theorem 6.1 (Conjugacy) *The posterior distribution of P given an i.i.d. sample $X_1, \dots, X_n$ from a $\mathrm{DP}(\alpha)$ process is $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$.*

Proof Because the Dirichlet process is tail free for any sequence of partitions by Theorem 4.9, and a given measurable partition $\{A_1, \dots, A_k\}$ of $\mathfrak{X}$ can be viewed as part of a sequence of successive binary partitions, the posterior distribution of the random vector $\big(P(A_1), \dots, P(A_k)\big)$ given $X_1, \dots, X_n$ is the same as the posterior distribution of this vector given the vector $N = (N_1, \dots, N_k)$ of cell counts, defined by $N_j = \#(1 \le i \le n{:}\, X_i \in A_j)$. Given P the vector N possesses a multinomial distribution with parameter $\big(P(A_1), \dots, P(A_k)\big)$, which has a $\mathrm{Dir}\big(k; \alpha(A_1), \dots, \alpha(A_k)\big)$ prior distribution. The posterior distribution can be obtained using Bayes' rule applied to these finite-dimensional vectors, as in Proposition 4.8. □

Theorem 6.1 can be remembered as the updating rule $\alpha \mapsto \alpha + \sum_{i=1}^n \delta_{X_i}$ for the base measure of the Dirichlet distribution. In terms of the parameterization $\alpha \leftrightarrow \big(M = |\alpha|, \bar\alpha\big)$ of the base measure, this rule takes the form

$$M \mapsto M + n \quad \text{ and } \quad \bar\alpha \mapsto \frac{M}{M+n}\bar\alpha + \frac{n}{M+n}\mathbb{P}_n, \tag{6.1}$$

where $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the *empirical distribution* of $X_1, \dots, X_n$. Because the mean measure of a Dirichlet process is the normalized base measure, we see that

$$\mathrm{E}\big(P(A) |\, X_1, \dots, X_n\big) = \frac{|\alpha|}{|\alpha|+n}\bar\alpha(A) + \frac{n}{|\alpha|+n}\mathbb{P}_n(A). \tag{6.2}$$

Thus the posterior mean (the "Bayes estimator" of P) is a convex combination of the prior

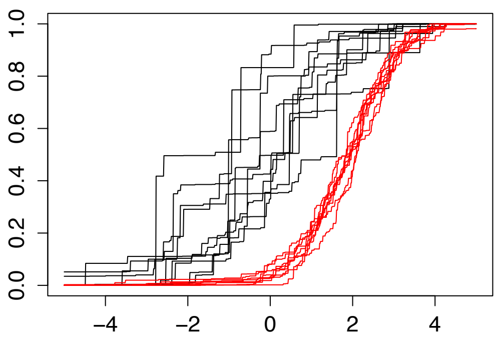

Figure 6.1 Cumulative distribution functions of 10 draws (black) from the Dirichlet process with base measure $5N(0,2)$, and of 10 draws (red) from the realization of the posterior distribution based on a sample of size 100 from a $N(2,1)$ distribution.

mean $\bar{\alpha}$ and the empirical distribution, with weights $M/(M+n)$ and $n/(M+n)$, respectively. For a given sample it is close to the prior mean if M is large, and close to the empirical distribution (which is based only on the data) if M is small. Thus M determines the extent to which the prior controls the posterior mean — a Dirichlet process prior with precision M contributes information equivalent to a sample of size M (although M is not restricted to integer values). This invites to view M as the *prior sample size*, or the "number of pre-experiment samples". In this interpretation the sum $M+n$ is the "posterior sample size".

For a fixed prior (i.e. fixed M) the posterior mean (6.2) behaves asymptotically as $n \to \infty$ like the empirical distribution $\mathbb{P}_n$ to the order $O(n^{-1})$, a.s.. Thus it possesses the same asymptotic properties as the empirical distribution. In particular, if $X_1, X_2, \ldots$ are sampled from a "true distribution"P_0, then the posterior mean will tend a.s. to P_0.

In addition the full posterior distribution will contract to its mean, whenever the posterior sample size tends to infinity. Indeed, by combining Theorem 6.1 and the formula for the variance of a Dirichlet variable, we see, for $\tilde{\mathbb{P}}_n$ the posterior mean (6.2),

$$\mathrm{var}\big(P(A)|\, X_1, \ldots, X_n\big) = \frac{\tilde{\mathbb{P}}_n(A)\tilde{\mathbb{P}}_n(A^c)}{1+M+n} \le \frac{1}{4(1+M+n)}. \qquad (6.3)$$

Consequently, if the data are sampled from a true distribution P_0, then the posterior distribution of P converges weakly to the measure degenerate at P_0. Formally, we can can state this as follows.

Corollary 6.2 *For any set A the posterior distribution of $P(A)$ given a random sample $X_1, \ldots, X_n$ of size n from a Dirichlet process tends in distribution to $\delta_{P_0(A)}$ as $n \to \infty$ for a.e. sequence $X_1, X_2, \ldots$ generated independently from a given distribution P_0.*

6.2 Predictive distribution

The joint distribution of a sequence $X_1, X_2, \dots$ generated from a Dirichlet process, has a complicated structure, but can be conveniently described by its sequence of *predictive distributions*: the laws of X_1, $X_2|\, X_1$, $X_3|\, X_1, X_2$, etc.

Because $\Pr(X_1 \in A) = \mathrm{E}\Pr(X_1 \in A|\, P) = \mathrm{E}P(A) = \bar{\alpha}(A)$, the marginal distribution of X_1 is $\bar{\alpha}$.

Because $X_2|\,(P, X_1) \sim P$ and $P|\, X_1 \sim \mathrm{DP}(\alpha+\delta_{X_1})$, we can apply the same reasoning again, but now conditionally given X_1, to see that $X_2|\, X_1$ follows the normalization of $\alpha + \delta_{X_1}$. This is a mixture of α and δ_{X_1} with weights $|\alpha|/(|\alpha|+1)$ and $1/(|\alpha|+1)$, respectively.

Repeating this argument, using that $P|\, X_1, \dots, X_{i-1} \sim \mathrm{DP}(\alpha + \sum_{j=1}^{i-1} \delta_{X_j})$, we find that

$$X_i|\, X_1, \dots, X_{i-1} \sim \begin{cases} \delta_{X_1}, & \text{with probability } \frac{1}{|\alpha|+i-1}, \\ \vdots & \qquad \vdots \\ \delta_{X_{i-1}}, & \text{with probability } \frac{1}{|\alpha|+i-1}, \\ \bar{\alpha}, & \text{with probability } \frac{|\alpha|}{|\alpha|+i-1}. \end{cases} \tag{6.4}$$

Being a mixture of a product of identical distributions, the joint distribution of $X_1, X_2, \dots$ is exchangeable, so re-labeling does not affect the structure of (6.4).

The recipe (6.4) is called the *generalized Polya urn scheme*, and can be viewed as a continuous analog of the familiar Polya urn scheme. Consider balls which can carry a continuum $\mathfrak{X}$ of "colors". Initially the "number of balls" is $M = |\alpha|$, which may be any positive number, and the colors are distributed according to $\bar{\alpha}$. We draw a ball from the collection, observe its color X_1, and return it to the urn along with an additional ball of the same color. The total number of balls is now $M + 1$, and the colors are distributed according to $(M\bar{\alpha} + \delta_{X_1})/(M + 1)$. We draw a ball from this updated urn, observe its color X_2, and return it to the urn along with an additional ball of the same color. The probability of picking up the ball that was added after the first draw is $1/(M + 1)$, in which case $X_2 = X_1$; otherwise, with probability $M/(M + 1)$, we make a fresh draw from the original urn. This process continues indefinitely, leading to the conditional distributions in (6.4).

6.3 Number of distinct values

It is clear from the preceding description that a realization of $(X_1, \dots, X_n)$ will have ties (equal values) with positive probability. For instance, with probability at least

$$\frac{1}{M+1}\frac{2}{M+2}\cdots\frac{n-1}{M+n-1}$$

all X_i will even be identical. For simplicity assume that the base measure α is non-atomic, so that the ith value X_i in the Polya scheme (6.4) is different from the previous $X_1, \dots, X_{i-1}$ if it is drawn from $\bar{\alpha}$. The vector $(X_1, \dots, X_n)$ then induces a random partition $\{\mathcal{P}_1, \dots, \mathcal{P}_{K_n}\}$ of the set of indices $\{1, 2, \dots, n\}$, corresponding to the ties, and given this partition the K_n distinct values are an i.i.d. sample from $\bar{\alpha}$.

The number of distinct values is remarkably small.

For $i \in \mathbb{N}$ define $D_i = 1$ if the ith observation X_i is a "new value", i.e. if $X_i \notin \{X_1, \ldots, X_{i-1}\}$, and set $D_i = 0$ otherwise. Then $K_n = \sum_{i=1}^n D_i$ is the number of distinct values among the first n observations.

Proposition 6.3 *If the base measure α is nonatomic and of strength $|\alpha| = M$, then the variables $D_1, D_2, \ldots$ are independent Bernoulli variables with success probabilities* $\Pr(D_i = 1) = M/(M + i - 1)$. *Consequently, for fixed M, as $n \to \infty$,*

(i) $\mathrm{E}(K_n) \asymp M \log n \asymp \operatorname{var}(K_n)$.
(ii) $K_n/\log n \to M$, *a.s.*
(iii) $(K_n - \mathrm{E}K_n)/\operatorname{sd}(K_n) \to_d \operatorname{Nor}(0, 1)$.

Proof The first assertion follows, because given $X_1, \ldots, X_{i-1}$ the variable X_i is "new" if and only if it is drawn from $\bar{\alpha}$, which happens with probability $M/(M + i - 1)$. Then assertion (i) can be derived from the exact formulas

$$\mathrm{E}(K_n) = \sum_{i=1}^n \frac{M}{M + i - 1}, \qquad \operatorname{var}(K_n) = \sum_{i=1}^n \frac{M(i-1)}{(M + i - 1)^2}.$$

Furthermore, assertion (ii) follows from Kolmogorov's strong law of large numbers for independent variables, since

$$\sum_{i=1}^\infty \frac{\operatorname{var}(D_i)}{(\log i)^2} = \sum_{i=1}^\infty \frac{M(i-1)}{(M + i - 1)^2 (\log i)^2} < \infty.$$

Next (iii) is a consequence of the Lindeberg central limit theorem. □

Thus the number of distinct values in a (large) sample from a distribution taken from a fixed Dirichlet prior is logarithmic in the sample size. Furthermore, the fluctuations of this number around its mean are of the order $\sqrt{\log n}$.

The following proposition gives the distribution of the partition $\{\mathcal{P}_1, \ldots, \mathcal{P}_{K_n}\}$ induced by $(X_1, \ldots, X_n)$. (This can be more formally defined as the equivalence classes under the relation $i \equiv j$ iff $X_i = X_j$.)

Proposition 6.4 *A random sample $X_1, \ldots, X_n$ from a Dirichlet process with nonatomic base measure of strength $|\alpha| = M$ induces a given partition of $\{1, 2, \ldots, n\}$ into k sets of sizes $n_1, \ldots, n_k$ with probability equal to*

$$\frac{M^k \Gamma(M) \prod_{j=1}^k \Gamma(n_j)}{\Gamma(M + n)}. \tag{6.5}$$

Proof By exchangeability the probability depends on the sizes of the partitioning sets only. The probability that the partitioning set of size n_1 consists of the first n_1 variables, the one of size n_2 of the next n_2 variables, etc. can be obtained by multiplying the appropriate conditional probabilities for the consecutive draws in the Polya urn scheme in their natural order of occurrence. For $r_j = \sum_{l=1}^j n_l$, it is given by

$$\frac{M}{M}\frac{1}{M+1}\frac{2}{M+2}\cdots\frac{n_1 - 1}{M + n_1 - 1}\frac{M}{M + n_1}\frac{1}{M + n_1 + 1} \times \cdots$$
$$\cdots \times \frac{M}{M + r_{k-1}}\frac{1}{M + r_{k-1} + 1}\cdots\frac{n_k - 1}{M + r_{k-1} + n_k - 1}.$$

This can be rewritten as in the proposition. □

6.4 Mixtures of Dirichlet processes

Application of the Dirichlet prior requires a choice of a base measure α. It is often reasonable to choose the center measure $\bar\alpha$ from a specific family such as the normal family, but then the parameters of the family must still be specified. It is natural to give these a further prior. Similarly, one may put a prior on the precision parameter $|\alpha|$.

For a base measure α_ξ that depends on a parameter ξ the Bayesian model then consists of the hierarchy

$$X_1,\ldots,X_n|\,P,\xi \overset{\text{iid}}{\sim} \sim P, \qquad P|\,\xi\sim \mathrm{DP}(\alpha_\xi), \qquad \xi\sim\pi. \tag{6.6}$$

We denote the induced (marginal) prior on P by $\mathrm{MDP}(\alpha_\xi,\xi\sim\pi)$. Many properties of this *mixture Dirichlet prior* follow immediately from those of a Dirichlet process. For instance, any P following an MDP is almost surely discrete. However, unlike a Dirichlet process, an MDP is not tail free.

Given ξ we can use the posterior updating rule for the ordinary Dirichlet process, and obtain that

$$P|\,\xi,X_1,\ldots,X_n\sim \mathrm{DP}(\alpha_\xi+n\mathbb{P}_n).$$

To obtain the posterior distribution of P given $X_1,\ldots,X_n$, we need to mix this over ξ relative to its posterior distribution given $X_1,\ldots,X_n$. By Bayes's theorem the latter has density proportional to

$$\xi\mapsto \pi(\xi)\,p(X_1,\ldots,X_n|\,\xi). \tag{6.7}$$

Here the marginal density of $X_1,\ldots,X_n$ given ξ (the second factor) is described by the generalized Polya urn scheme (6.4) with α_ξ instead of α. In general, this has a somewhat complicated structure due to the ties between the observations. However, for a posterior calculation we condition on the observed data $X_1,\ldots,X_n$, and know the partition that they generate. Given this information the density takes a simple form. For instance, if the observations are distinct (which happens with probability one if the observations actually follow a continuous distribution), then the Polya urn scheme must have simply generated a random sample from the normalized base measure $\bar\alpha_\xi$, in which case the preceding display becomes

$$\pi(\xi)\prod_{i=1}^n d\alpha_\xi(X_i)\prod_{i=1}^n\frac{1}{|\alpha_\xi|+i-1},$$

for $d\alpha_{xi}$ a density of α_ξ. Further calculations depend on the specific family and its parameterization.

Typically the precision parameter M and center measure G in $\alpha=MG$ will be modelled as independent under the prior. The posterior calculation then factorizes in these two parameters. To see this, consider the following scheme to generate the parameters and observations:

(i) Generate M from its prior.

(ii) Given M generate a random partition $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_{K_n}\}$ according to the distribution given in Proposition 6.4.
(iii) Generate G from its prior, independently of $(M, \mathcal{P})$.
(iv) Given $(\mathcal{P}, G)$ generate a random sample of size K_n from G, independently of M, and set X_i with $i \in \mathcal{P}_j$ equal to the jth value in this sample.

By the description of the Polya urn scheme this indeed gives a sample $X_1, \ldots, X_n$ from the mixture of Dirichlet processes $\mathrm{MDP}(MG, M \sim \pi, G \sim \pi)$. We may now formally write the density of $(M, \mathcal{P}, G, X_1, \ldots, X_n)$ in the form, with π abusively denoting prior densities for both M and G and p conditional densities of observed quantities,

$$\pi(M)\, p(\mathcal{P}|\, M)\, \pi(G)\, p(X_1, \ldots, X_n|\, G, \mathcal{P}).$$

Since this factorizes in terms involving M and G, these parameters are also independent under the posterior distribution, and the computation of their posterior distributions can be separated.

The term involving M depends on the data through K_n only (the latter variable is *sufficient* for M). Indeed, by Proposition 6.4 it is proportional to,

$$M \mapsto \pi(M) \frac{M^{K_n}\Gamma(M)}{\Gamma(M+n)} \propto \pi(M) M^{K_n} \int_0^1 \eta^{M-1}(1-\eta)^{n-1}\, d\eta.$$

Rather than by (numerically) integrating this expression, the posterior density is typically computed by simulation. xsSuppose that $M \sim \mathrm{Ga}(a, b)$ a priori, and consider a fictitious random vector (M, η) with $0 \leq \eta \leq 1$ and joint (Lebesgue) density proportional to

$$\pi(M) M^{K_n} \eta^{M-1}(1-\eta)^{n-1} \propto M^{a+K_n-1} e^{-M(b-\log \eta)} \eta^{-1}(1-\eta)^{n-1}.$$

Then by the preceding display the marginal density of M is equal to its posterior density (given K_n, which is fixed for the calculation). Thus simulating from the distribution of (M, η) and dropping η simulates M from its posterior distribution. The conditional distributions are given by

$$M|\, \eta, K_n \sim \mathrm{Ga}(a + K_n, b - \log \eta), \qquad \eta|\, M, K_n \sim \mathrm{Be}(M, n). \tag{6.8}$$

We can use these in a *Gibbs sampling scheme*: given an arbitrary starting value η_0 we generate a sequence $M_1, \eta_1, M_2, \eta_2, M_3, \ldots$, by repeatedly generating M from its conditional distribution given (η, K_n) and η from its conditional distribution given (M, K_n), each time setting the conditioning variable (η or M) equal to its last value. After an initial *burn-in* the values $M_k, M_{k+1}, \ldots$ will be approximately from the posterior distribution of M given K_n.

6.5 Dirichlet process mixtures

Because the Dirichlet process is discrete, it is a useless prior when we wish to estimate a density. This can be remedied by convolving it with a kernel. For each θ in a parameter set Θ let $x \mapsto \psi(x, \theta)$ be a probability density function, measurable in its two arguments. For a measure F on Θ define a *mixture density* by

$$p_F(x) = \int \psi(x, \theta)\, dF(\theta).$$

By equipping F with a prior, we obtain a prior on densities. Densities p_F with F following a Dirichlet process prior are known as *Dirichlet mixtures*. If the kernel also depends on an additional parameter $\varphi \in \Phi$, giving mixtures $p_{F,\varphi}(x) = \int \psi(x, \theta, \varphi)\, dF(\theta)$, it is more appropriate to call the result a "mixture of Dirichlet mixture", but the nomenclature Dirichlet mixture even for this case seems more convenient.

In this section we discuss methods of posterior computation for these mixtures. For $x \mapsto \psi(x; \theta, \varphi)$ a probability density function (relative to a given σ-finite dominating measure ν), consider

$$X_i \overset{\text{iid}}{\sim} p_{F,\varphi}(x) = \int \psi(x; \theta, \varphi)\, dF(\theta), \qquad i = 1, \ldots, n, \tag{6.9}$$

We equip F and φ with independent priors $F \sim \mathrm{DP}(\alpha)$ and $\varphi \sim \pi$. The resulting model can be equivalently written in terms of n *latent variables* $\theta_1, \ldots, \theta_n$ as

$$X_i |\, \theta_i, \varphi, F \overset{\text{ind}}{\sim} \psi(\cdot; \theta_i, \varphi), \qquad \theta_i |\, F, \varphi \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha), \qquad \varphi \sim \pi. \tag{6.10}$$

The posterior distribution of any object of interest can be described in terms of the posterior distribution of (F, φ) given $X_1, \ldots, X_n$. The latent variables $\theta_1, \ldots, \theta_n$ help to make the description simpler, since $F |\, \theta_1, \ldots, \theta_n \sim \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$, and given $\theta_1, \ldots, \theta_n$, the observations $X_1, \ldots, X_n$ and F are independent. Hence the conditional distribution of F given $\theta_1, \ldots, \theta_n, X_1, \ldots, X_n$ is free of the observations. In particular, for any measurable function ψ, in view of Exercise 4.1,

$$\mathrm{E}\Bigl(\int \psi\, dF |\, \varphi, \theta_1, \ldots, \theta_n, X_1, \ldots, X_n\Bigr) = \frac{1}{|\alpha| + n}\Bigl[\int \psi\, d\alpha + \sum_{j=1}^n \psi(\theta_j)\Bigr]. \tag{6.11}$$

The advantage of this representation is that the infinite-dimensional parameter F has been eliminated. To compute the posterior expectation it now suffies to average out the right hand side of (6.11) with respect to the posterior distribution of $(\theta_1, \ldots, \theta_n)$, and that of φ.

Example 6.5 (Density estimation) The choice $\psi(\theta) = \psi(x, \theta, \varphi)$ in (6.11) gives the density $\int \psi(x, \theta, \varphi)\, dF(\theta) = p_{F,\varphi}(x)$. Thus the posterior mean density satisfies

$$\mathrm{E}\bigl(p_{F,\varphi}(x) |\, \varphi, X_1, \ldots, X_n\bigr) = \frac{1}{|\alpha| + n}\Bigl[\int \psi(x; \theta, \varphi)\, d\alpha(\theta) + \mathrm{E}\Bigl(\sum_{j=1}^n \psi(x; \theta_j, \varphi) |\, X_1, \ldots, X_n\Bigr)\Bigr].$$

This consists of a part attributable to the prior and a part due to observations. In practice the latter is computed by simulating many samples $(\theta_1 \ldots, \theta_n)$ from its posterior distribution.

Analytical formulas for the posterior distribution corresponding to a Dirichlet mixture are possible, but too unwieldy for practical implementation. Computation is typically done by simulation. The next theorem explains a *Gibbs sampling scheme* to simulate from the posterior distribution of $(\theta_1, \ldots, \theta_n)$, based on a weighted generalized Polya urn scheme. Inclusion of a possible parameter φ and other hyperparameters is tackled in the next section.

A *Gibbs sampler* in general is a method for simulating from the joint distribution of a number of variables. It simply updates the variables one-by-one by simulating a new variable from its conditional distribution given the other variables. By repeating this indefinitely a sequence of vectors is created that after an initial "burn-in period" can be viewed as sampled from the target distribution. More precisely, the sequence of vectors forms a Markov chain with the target distribution as its stationary distribution.

We use the subscript $-i$ to denote every index $j \neq i$, and $\theta_{-i} = (\theta_j : j \neq i)$.

Theorem 6.6 (Gibbs sampler) *The conditional posterior distribution of θ_i is given by:*

$$\theta_i |\, \theta_{-i}, \varphi, X_1, \ldots, X_n \sim \sum_{j \neq i} q_{i,j} \delta_{\theta_j} + q_{i,0} G_{b,i}, \tag{6.12}$$

where $(q_{i,j} : j \in \{0, 1, \ldots, n\} - \{i\})$ is the probability vector satisfying

$$q_{i,j} \propto \begin{cases} \psi(X_i; \theta_j, \varphi), & j \neq i, j \geq 1, \\ \int \psi(X_i; \theta, \varphi)\, d\alpha(\theta), & j = 0, \end{cases} \tag{6.13}$$

and $G_{b,i}$ is the "baseline posterior measure" given by

$$dG_{b,i}(\theta |\, \varphi, X_i) \propto \psi(X_i; \theta, \varphi)\, d\alpha(\theta). \tag{6.14}$$

Proof Since the parameter φ is fixed throughout, we suppress it from the notation. For measurable sets A and B,

$$\mathrm{E}\big(1\!\!1_A(X_i) 1\!\!1_B(\theta_i) |\, \theta_{-i}, X_{-i}\big) = \mathrm{E}\Big(\mathrm{E}\big(1\!\!1_A(X_i) 1\!\!1_B(\theta_i) |\, F, \theta_{-i}, X_{-i}\big) |\, \theta_{-i}, X_{-i}\Big).$$

Because (θ_i, X_i) is conditionally independent of (θ_{-i}, X_{-i}) given F, the inner conditional expectation is equal to $\mathrm{E}\big(1\!\!1_A(X_i) 1\!\!1_B(\theta_i) |\, F\big) = \int\int 1\!\!1_A(x) 1\!\!1_B(\theta) \psi(x; \theta)\, d\mu(x)\, dF(\theta)$. In the outer layer of conditioning the variables X_{-i} are superfluous, by the conditional independence of F and X_{-i} given θ_{-i}. Therefore, by Exercise 4.1 the preceding display is equal to

$$\frac{1}{|\alpha| + n} \int\int 1\!\!1_A(x) 1\!\!1_B(\theta) \psi(x; \theta)\, d\mu(x)\, d\Big(\alpha + \sum_{j \neq i} \delta_{\theta_j}\Big)(\theta).$$

This determines the joint conditional distribution of (X_i, θ_i) given (θ_{-i}, X_{-i}). By Bayes's rule (applied to this joint law conditionally given (θ_{-i}, X_{-i})) we infer that

$$\Pr\big(\theta_i \in B |\, X_i, \theta_{-i}, X_{-i}\big) = \frac{\int_B \psi(X_i; \theta)\, d(\alpha + \sum_{j \neq i} \delta_{\theta_j})(\theta)}{\int \psi(X_i; \theta)\, d(\alpha + \sum_{j \neq i} \delta_{\theta_j})(\theta)}.$$

This in turn is equivalent to the assertion of the theorem. □

6.5.1 MCMC method

In this section we present an algorithm to simulate from the posterior distribution in the MDP model:

$$X_i |\, \theta_i, \varphi, M, \xi, F \overset{\text{ind}}{\sim} \psi(\cdot; \theta_i, \varphi), \quad \theta_i |\, F, \varphi, M, \xi \overset{\text{iid}}{\sim} F, \quad F |\, M, \xi \sim \mathrm{DP}(M, G_\xi),$$

where φ, M and ξ are independently generated hyperparameters. The basic algorithm uses the Gibbs sampling scheme of Theorem 6.6 to generate $\theta_1, \ldots, \theta_n$ given $X_1, \ldots, X_n$ in combination with the Gibbs sampler for the posterior distribution of M given in Section 6.4, and/or additional Gibbs steps. The prior densities of the hyperparameters are denoted by a generic π.

Algorithm Generate samples by sequentially executing steps (i)–(iv) below:

(i) Given the observations and φ, M and ξ, update each θ_i sequentially using (6.12) inside a loop $i = 1, \ldots, n$.

(ii) Update $\varphi \sim p(\varphi | \, \theta_1, \ldots, \theta_n, X_1, \ldots, X_n) \propto \pi(\varphi) \prod_{i=1}^n \psi(X_i; \theta_i, \varphi)$.

(iii) Update $\xi \sim p(\xi | \, \theta_1, \ldots, \theta_n) \propto \pi(\xi) p(\theta_1, \ldots, \theta_n | \, \xi)$, where the marginal distribution of $(\theta_1, \ldots, \theta_n)$ is as in the Polya scheme (6.4).

(iv) Update M and next the auxiliary variable η using (6.8), for K_n the number of distinct values in $\{\theta_1, \ldots, \theta_n\}$.

Exercises

6.1 Let ψ be a given bounded measurable function. Show that if $P \sim \mathrm{DP}(\alpha)$ and $X_1, \ldots, X_n | \, P \overset{\text{iid}}{\sim} P$, then the posterior distribution of $\int \psi \, dP$ given $X_1, \ldots, X_n$ tends in distribution to a Dirac measure at $\int \psi \, dP_0$ for a.e. sequence $X_1, X_2, \ldots$ generated iid from P_0.

6.2 In the model (6.6) assume that the total mass $|\alpha_\xi|$ is bounded uniformly in ξ. Show that the posterior distribution of $P(A)$ is consistent.

6.3 Simulate and plot the cumulative distribution functions of realizations of some posterior Dirichlet processes. First use several fixed prior strengths. Second put a Gamma prior on the prior strength.

7

Consistency

7.1 Definition

For every $n \in \mathbb{N}$ let $X^{(n)}$ be an observation in a sample space $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)})$ with distribution $P_\theta^{(n)}$ indexed by a parameter θ belonging to a metric space Θ. For instance $X^{(n)}$ may be sample of size n from a given distribution P_θ, and $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)})$ the corresponding product probabiity space. Given a prior Π on the Borel sets of Θ, let $\Pi_n(\cdot|\, X^{(n)})$ be a version of the posterior distribution.

Definition 7.1 (Posterior consistency) The posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is said to be *(weakly) consistent* at $\theta_0 \in \Theta$ if $\Pi_n\bigl(\theta\colon d(\theta,\theta_0) > \epsilon|\, X^{(n)}\bigr) \to 0$ in $P_{\theta_0}^{(n)}$-probability, as $n \to \infty$, for every $\epsilon > 0$. The posterior distribution is said to be *strongly consistent* at $\theta_0 \in \Theta$ if the convergence is in the almost sure sense.

Both forms of consistency are of interest. Naturally, strong consistency is more appealing as it is stronger, but it may require more assumptions. To begin with it presumes that the observations $X^{(n)}$ are defined on a common underlying probability space (with for each n the measure $P_\theta^{(n)}$ equal to the image $P_\theta^{(\infty)} \circ (X^{(n)})^{-1}$ of the probability measure $P_\theta^{(\infty)}$ on this space), or at least that their joint distribution is defined, whereas weak consistency makes perfect sense without any relation between the observations across n.

Consistency entails that the full posterior distribution contracts to within arbitrarily small distance ϵ to the true parameter θ_0. It can also be summarized as saying that the posterior distribution converge weakly to a Dirac measure at θ_0, in probability or almost surely.

Naturally an appropriate summary of its location should provide a point estimator that is consistent in the usual sense of consistency of estimators. The following proposition gives a summary that works without further conditions. (The value $1/2$ could be replaced by any other number between 0 and 1.)

Proposition 7.2 (Point estimator) *Suppose that the posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is consistent (or strongly consistent) at θ_0 relative to the metric d on Θ. Then $\hat\theta_n$ defined as the center of a (nearly) smallest ball that contains posterior mass at least $1/2$ satisfies $d(\hat\theta_n, \theta_0) \to 0$ in $P_{\theta_0}^{(n)}$-probability (or almost surely $[P_{\theta_0}^{(\infty)}]$, respectively).*

Proof For $B(\theta, r) = \{s \in \Theta\colon d(s,\theta) \le r\}$ the closed ball of radius r around $\theta \in \Theta$, let $\hat r_n(\theta) = \inf\{r\colon \Pi_n\bigl(B(\theta,r)|\, X^{(n)}\bigr) \ge 1/2\}$, where the infimum over the empty set is ∞. Taking the balls closed ensures that $\Pi_n\bigl(B(\theta, \hat r_n(\theta))|\, X^{(n)}\bigr) \ge 1/2$, for every θ. Let $\hat\theta_n$ be a near minimizer of $\theta \mapsto \hat r_n(\theta)$ in the sense that $\hat r_n(\hat\theta_n) \le \inf_\theta \hat r_n(\theta) + 1/n$.

By consistency $\Pi_n\bigl(B(\theta_0,\epsilon)|\,X^{(n)}\bigr) \to 1$ in probability or almost surely, for every $\epsilon > 0$. As a first consequence $\hat r_n(\theta_0) \le \epsilon$ with probability tending to one, or eventually almost surely, and hence $\hat r_n(\hat\theta_n) \le \hat r_n(\theta_0) + 1/n$ is bounded by $\epsilon + 1/n$ with probability tending to one, or eventually almost surely. As a second consequence the balls $B(\theta_0,\epsilon)$ and $B\bigl(\hat\theta_n, \hat r_n(\hat\theta_n)\bigr)$ cannot be disjoint, as their union would contain mass nearly $1 + 1/2$. This shows that $d(\theta_0,\hat\theta_n) \le \epsilon + \hat r_n(\hat\theta_n)$ with probability tending to one, or eventually almost surely, which is further bounded by $2\epsilon + 1/n$. □

An alternative point estimator is the *posterior mean* $\int \theta\, d\Pi_n(\theta|\,X^{(n)})$ (available when Θ has a vector space structure). This is attractive for computational reasons, as it can be approximated by the average of the output of a simulation run. Usually the posterior mean is also consistent, but in general this may require additional assumptions. For instance, weak convergence to a Dirac measure on a Euclidean space does not imply convergence of moments.

7.2 Doob's theorem

Doob's theorem basically says that for any fixed prior, the posterior distribution is consistent at every θ except those in a "bad set" that is "small" when seen from the prior point of view. We first present the theorem, and next argue that the message is not as positive as it may first seem. Because then it is not as useful after all, we state only the result for i.i.d. observations and omit the proof. (See e.g. the book Asymptotic Statistics, Chapter 10, by van der Vaart.)

Theorem 7.3 (Doob) *Let $(\mathfrak{X}, \mathscr{X}, P_\theta{:}\, \theta \in \Theta)$ be experiments with $(\mathfrak{X}, \mathscr{X})$ a standard Borel space and Θ a Borel subset of a Polish space such that $\theta \mapsto P_\theta(A)$ is Borel measurable for every $A \in \mathscr{X}$ and the map $\theta \mapsto P_\theta$ is one-to-one. Then for any prior Π on the Borel sets of Θ the posterior $\Pi_n(\cdot|\,X_1,\ldots,X_n)$ in the model $X_1,\ldots,X_n|\,\theta \overset{iid}{\sim} p_\theta$ and $\theta \sim \Pi$ is strongly consistent at θ, for Π-almost every θ.*

Doob's theorem is remarkable in many respects. Virtually no condition is imposed on the model or the parameter space, but nevertheless a Bayesian will "almost always" have consistency, as long as she is certain of her prior. Since null sets are negligibly small, a troublesome value of the parameter "will not obtain".

However, such a view is very dogmatic. No one in practice can be certain of the prior, and troublesome values of the parameter may really obtain. In fact, the Π-null set could be very large if *not* judged from the point of view of the prior. To see an extreme example, consider a prior that assigns all its mass to some fixed point θ_0. The posterior then also assigns mass one to θ_0 and hence is *in*consistent at every $\theta \neq \theta_0$. Doob's theorem is still true, of course; the point is that the set $\{\theta{:}\, \theta \neq \theta_0\}$ is a null set under the present prior. Thus Doob's theorem should not create a false sense of satisfaction about Bayesian procedures in general. It is important to know, for a given "reasonable" prior, at which parameter values consistency holds. Consistency at every parameter in a set of prior probability one is not enough.

An exception is the case that the parameter set Θ is countable. Then Doob's theorem shows that consistency holds at θ as long as Π assigns positive mass to it. More generally, consistency holds at any atom of a prior. However, even in these cases the theorem is of "asymptopia" type only, in that at best it gives convergence without quantification of the approximation error, or uniformity in the parameter.

7.3 Schwartz's theorem and its extensions

In this section we take the parameter equal to a density, relative to a given dominating measure ν on the sample space $(\mathfrak{X}, \mathscr{X})$. We denote this parameter by p rather than θ, and the corresponding parameter set by $\mathcal{P}$. We consider estimating p based on a random sample $X_1, \ldots, X_n$ of observations, with true density p_0. As notational convention we denote a density by a lower case letter p and the measure induced by it by the uppercase letter P. The parameter set is equipped with a metric that is unspecified for the moment.

A key condition for posterior consistency is that the prior assigns positive probability to any Kullback-Leibler (or KL) neighborhood of the true density. The *Kullback-Leibler divergence* between two densities p_0 and p is defined as

$$K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu.$$

Note that it is asymmetric in its arguments. For a set $\mathcal{P}_0$ of densities we write $K(p_0; \mathcal{P}_0) = \inf_{p\in\mathcal{P}_0} K(p_0; p)$ for the minimum divergence.

Definition 7.4 A density p_0 is said to possess the *Kullback-Leibler property* relative to a prior Π if $\Pi\big(p\colon K(p_0; p) < \epsilon\big) > 0$ for every $\epsilon > 0$. This is denoted $p_0 \in \mathrm{KL}(\Pi)$. Alternatively, we say that p_0 belongs to the *Kullback-Leibler support* of Π.[1]

Schwartz's theorem is the basic result on posterior consistency for dominated models. It has two conditions: the true density p_0 should be in the KL-support of the prior, and the hypothesis $p = p_0$ should be testable against complements of neighborhoods of p_0. The first is clearly a Bayesian condition, but the second may be considered a condition to enable recovery of p_0 by any statistical method. Although in its original form the theorem has limited applicability, extensions go far deeper, and lead to a rich theory of posterior consistency. Also the theory of convergence rates, developed in Chapter 9, uses similar arguments.

In the present context *tests* ϕ_n are understood to refer both to measurable mappings $\phi_n\colon \mathfrak{X}^n \to [0,1]$, and to the corresponding statistics $\phi_n(X_1, \ldots, X_n)$. The interpretation of a test ϕ_n is that a null hypothesis is rejected with probability ϕ_n, whence $P^n\phi_n$ is the probability of rejection if the data are sampled from P. It follows that $P_0^n\phi_n$ is the probability of a type I error for testing $H_0\colon P = P_0$, and $P^n(1-\phi_n)$ is the probability of a type II error if $P \neq P_0$.

Theorem 7.5 (Schwartz) *If* $p_0 \in \mathrm{KL}(\Pi)$ *and for every neighbourhood* $\mathcal{U}$ *of* p_0 *there exist tests* ϕ_n *such that* $P_0^n\phi_n \to 0$ *and* $\sup_{p\in\mathcal{U}^c} P^n(1-\phi_n) \to 0$, *then the posterior distribution* $\Pi_n(\cdot|\, X_1, \ldots, X_n)$ *in the model* $X_1 \ldots, X_n|\, p \overset{iid}{\sim} p$ *and* $p \sim \Pi$ *is strongly consistent at* p_0.

Proof It must be shown that $\Pi_n(\mathcal{U}^c|\, X_1, \ldots, X_n) \to 0$ almost surely, for every given neighborhood $\mathcal{U}$ of p_0. We shall show that it is not a loss of generality to assume that the

[1] The Kullback-Leibler divergence is typically measurable in its second argument, and then Kullback-Leibler neighborhoods are measurable in the space of densities. If not, then we interpret the KL-property in the sense of inner probability: it suffices that there exists measurable sets $\mathcal{B} \subset \big\{p\colon K(p_0; p) < \epsilon\big\}$ with $\Pi(\mathcal{B}) > 0$.

tests ϕ_n as in the theorem have exponentially small error probabilities in the sense that, for some positive constant C,

$$P_0^n\phi_n \leq e^{-Cn}, \qquad \sup_{p\in\mathcal{U}^c} P^n(1-\phi_n) \leq e^{-Cn}.$$

Then the theorem follows from an application of Theorem 7.8 below, with $\mathcal{P}_n = \mathcal{P}$ for every n.

There exists n_0 such that $P_0^{n_0}\phi_{n_0} < 1/4 < 3/4 < Q^{n_0}\phi_{n_0}$, for every $Q \in \mathcal{U}^c$. For a given n divide the observations in $l{:}= \lfloor n/n_0 \rfloor \asymp n$ groups of size n_0, and a remainder. The variables $Y_1, \ldots, Y_l$ obtained by applying the test ϕ_{n_0} to these n_0 groups are independent with mean smaller than $1/4$ under P and bigger than $3/4$ under every Q. Consider the test that rejects if their average $\bar{Y}_{n_0}$ is bigger than $1/2$. Because $0 \leq Y_i \leq 1$, we can apply Hoeffding's inequality, Lemma 7.10, to see that this has error probabilities of the desired type, first to the events $\{\bar{Y}_n > 1/2\} \subset \{\bar{Y}_n - \mathrm{E}_0\bar{Y}_n > 1/4\}$ under P_0^n, and a second time to the events $\{\bar{Y}_n < 1/2\} \subset \{\bar{Y}_n - \mathrm{E}_Q\bar{Y}_n < -1/4\}$ under Q^n. □

Example 7.6 (Finite-dimensional models) If the model is smoothly parameterized by a finite-dimensional parameter that varies over a bounded set, then consistent tests as required in Schwartz's theorem, Theorem 7.5, exist under mere regularity conditions on the model. For unbounded Euclidean sets some minor conditions are needed.

Example 7.7 (Consistency for weak topology) The weak topology on the set of probability measures can also be viewed as a topology on the corresponding densities $\mathcal{P}$. For this topology consistent tests as in Schwartz's theorem, Theorem 7.5, always exist. Therefore, the posterior distribution is consistent for the weak topology at any density p_0 that has the Kullback-Leibler property for the prior.

To construct the tests observe that finite intersections of sets of the type $\mathcal{U} = \{p{:}\, P\psi < P_0\psi + \epsilon\}$ form a base for the weak neighborhood system at a probability measure P_0, when ψ varies over the continuous functions $\psi{:}\, \mathfrak{X} \to [0,1]$ and $\epsilon > 0$ (see Lemma 7.11). Given a test for any neighbourhood of this type, we can form a test for finite intersections by rejecting P_0 as soon as P_0 is rejected for one of the finitely many neighbourhoods. The resulting error probabilities are bounded by the sum of the error probabilities of the finitely many tests, and hence will tend to zero.

Now by Hoeffding's inequality, Lemma 7.10, for $\psi{:}\, \mathfrak{X} \to [0,1]$ the test

$$\phi_n = 1\!\!1\Big\{\frac{1}{n}\sum_{i=1}^n \psi(X_i) > P_0\psi + \epsilon/2\Big\}$$

has type I error satisfying $P_0^n\phi_n \leq e^{-n\epsilon^2/2}$. Furthermore, since $P_0\psi - P\psi < -\epsilon$ whenever $P \in \mathcal{U}^c$, we have $P^n(1-\phi_n) \leq P^n\big(n^{-1}\sum_{i=1}^n(\psi(X_i) - P\psi) < -\epsilon/2\big)$ for $P \in \mathcal{U}^c$ and this is bounded by $e^{-n\epsilon^2/2}$, by a second application of Hoeffding's inequality.

In its original form Schwartz's theorem requires that the complement of every neighbourhood of p_0 can be "tested away". For strong metrics, such as the L_1-distance, such tests may not exist, even though the posterior distribution may be consistent. The following extension of the theorem is useful for these situations. The idea is that the posterior distribution will always give vanishing mass to sets of very small prior mass. Such sets need not be tested.

Theorem 7.8 (Extension of Schwartz's theorem) *If $p_0\in\mathrm{KL}(\Pi)$ and for every neighbourhood $\mathcal{U}$ of p_0 there exist a constant $C>0$, measurable sets $\mathcal{P}_n\subset\mathcal{P}$ and tests ϕ_n such that*

$$\Pi(\mathcal{P}-\mathcal{P}_n)<e^{-Cn},\qquad P_0^n\phi_n\le e^{-Cn},\qquad \sup_{p\in\mathcal{P}_n\cap\mathcal{U}^c}P^n(1-\phi_n)\le e^{-Cn},$$

then the posterior distribution $\Pi_n(\cdot|\,X_1,\ldots,X_n)$ in the model $X_1\ldots,X_n|\,p\overset{iid}{\sim}p$ and $p\sim\Pi$ is strongly consistent at p_0.

Proof We first show that for any $\epsilon>0$ eventually a.s. $[P_0^\infty]$:

$$\int\prod_{i=1}^n\frac{p}{p_0}(X_i)\,d\Pi(p)\ge\Pi\bigl(p\colon K(p_0;p)<\epsilon\bigr)e^{-n\epsilon}.\tag{7.1}$$

For any set $\mathcal{P}_0\subset\mathcal{P}$ the integral is bounded below by $\Pi(\mathcal{P}_0)\int\prod_{i=1}^n(p/p_0)(X_i)\,d\Pi_0(p)$, for Π_0 the renormalized restriction $\Pi(\cdot\cap\mathcal{P}_0)/\Pi(\mathcal{P}_0)$ to $\mathcal{P}_0$. Therefore the logarithm of the integral is bounded below by

$$\log\Pi(\mathcal{P}_0)+\log\int\prod_{i=1}^n\frac{p}{p_0}(X_i)\,d\Pi_0(p)\ge\log\Pi(\mathcal{P}_0)+\int\log\prod_{i=1}^n\frac{p}{p_0}(X_i)\,d\Pi_0(p),$$

by Jensen's inequality applied to the logarithm (which is concave). The second term is n times the average

$$\frac{1}{n}\sum_{i=1}^n\int\log\frac{p}{p_0}(X_i)\,d\Pi_0(p)\to P_0\int\log\frac{p}{p_0}\,d\Pi_0(p),\qquad a.s.$$

by the strong law of large numbers. The right side is $-\int K(p_0;p)\,d\Pi_0(p)$, and is strictly bigger than $-\epsilon$ for $\mathcal{P}_0=\{p\colon K(p_0;p)<\epsilon\}$. This implies (7.1).

Next fix a neighbourhood $\mathcal{U}$ of p_0, and let C, $\mathcal{P}_n$ and the tests ϕ_n be as in the statement of the theorem. We shall show separately that $\Pi_n(\mathcal{P}_n\cap\mathcal{U}^c|\,X_1,\ldots,X_n)\to0$ and that $\Pi_n(\mathcal{P}_n^c|\,X_1,\ldots,X_n)\to0$, almost surely.

In view of Bayes's rule (2.1),

$$\Pi_n(\mathcal{P}_n\cap\mathcal{U}^c|\,X_1,\ldots,X_n)\le\phi_n+\frac{(1-\phi_n)\int_{\mathcal{P}_n\cap\mathcal{U}^c}\prod_{i=1}^n(p/p_0)(X_i)\,d\Pi(p)}{\int\prod_{i=1}^n(p/p_0)(X_i)\,d\Pi(p)}.$$

The expectation of the first term is bounded by e^{-Cn} by assumption, whence $\sum_n P_0^n(\phi_n>\delta)<\sum_n\delta^{-1}e^{-Cn}<\infty$, by Markov's inequality. This implies that $\phi_n\to0$ almost surely, by the Borel-Cantelli lemma.

By (7.1) and the fact that p_0 is in the Kullback-Leibler support of Π the denominator of the second term is bounded below by a constant times $e^{-n\epsilon}$ eventually a.s., for every given ϵ. Thus the left side of the display tends to zero if $e^{n\epsilon}$ times the numerator tends to zero. By Fubini's theorem,

$$\begin{aligned}P_0^n\Bigl((1-\phi_n)\int_{\mathcal{P}_n\cap\mathcal{U}}\prod_{i=1}^n\frac{p}{p_0}(X_i)\,d\Pi(p)\Bigr)&=\int_{\mathcal{P}_n\cap\mathcal{U}^c}P_0^n\Bigl[(1-\phi_n)\prod_{i=1}^n\frac{p}{p_0}(X_i)\Bigr]\,d\Pi(p)\\&\le\int_{\mathcal{P}_n\cap\mathcal{U}^c}P^n(1-\phi_n)\,d\Pi(p)\le e^{-Cn}.\end{aligned}$$

Since $\sum_n e^{n\epsilon}e^{-Cn} < \infty$ if $\epsilon < C$, the desired convergence of $e^{n\epsilon}$ times the numerator follows by Markov's inequality.

Finally we apply the argument of the preceding paragraph with $\mathcal{P}_n \cap \mathcal{U}^c$ replaced by $\mathcal{P}_n^c$ and the tests $\phi_n = 0$ instead of the given tests. The "power" $P^n(1-\phi_n)$ of this test is equal to one, but the final term of the preceding display can be bounded by $\Pi(\mathcal{P}_n^c)$, which is also of the order e^{-Cn}, by assumption. This shows that $\Pi_n(\mathcal{P}_n^c |\, X_1, \ldots, X_n) \to 0$, almost surely. □

The construction of appropriate tests is deferred to a later chapter. For a strong metric, such as the L_1-distance, their existence is not automatic, but they do exist for models that are not too big. We close this section by stating a theorem in this direction; its proof will be derived later.

We write $N(\epsilon, \mathcal{P}, d)$ for the minimal number of d-balls of radius ϵ needed to cover a set $\mathcal{P}$. This is called the *covering number* of $\mathcal{P}$ and is discussed in Chapter 8.

Theorem 7.9 (Consistency in total variation) *The posterior distribution is strongly consistent relative to the L_1-distance at every $p_0 \in \mathrm{KL}(\Pi)$ if for every $\epsilon > 0$ there exist a partition $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ (which may depend on ϵ) such that, for constants $C > 0$, $\xi < 1/2$, and sufficiently large n,*

(i) $\Pi(\mathcal{P}_{n,2}) \leq e^{-Cn}$.
(ii) $\log N\big(\epsilon, \mathcal{P}_{n,1}, \|\cdot\|_1\big) \leq \xi n \epsilon^2$.

7.4 COMPLEMENTS

Lemma 7.10 (Hoeffding) *For any independent random variables $X_1, \ldots, X_n$ such that $a \leq X_i \leq b$ for every i, and any $t > 0$,*

$$\mathrm{P}\big(\bar{X}_n - \mathrm{E}\bar{X}_n \geq t\big) \leq e^{-2nt^2/(b-a)^2}.$$

Proof By Markov's inequality applied to the variable $e^{hn(\bar{X}_n - \mathrm{E}\bar{X}_n)}$, for $h > 0$ to be chosen later, we obtain

$$\mathrm{P}\left(\sum_{i=1}^n (X_i - \mathrm{E}(X_i)) \geq t\right) \leq e^{-hnt}\mathrm{E}\prod_{i=1}^n e^{h(X_i - \mathrm{E}(X_i))}.$$

By independence of the X_i the order of expectation and product on the right side can be swapped. By convexity of the exponential function $e^{hX} \leq ((b-X)e^{ha} + (X-a)e^{hb})/(b-a)$ whenever $a \leq X \leq b$, whence, by taking expectation,

$$\mathrm{E}(e^{hX}) \leq e^{ha}\frac{b - \mathrm{E}(X)}{b-a} + e^{hb}\frac{\mathrm{E}(X) - a}{b-a} = e^{g(\xi)},$$

where $g(\xi) = \log(1 - p + pe^{\xi}) - p\xi$, for $\xi = (b-a)h$ and $p = (\mathrm{E}(X) - a)/(b-a)$.

Now $g(0) = 0$, $g'(0) = 0$ and $g''(\xi) = (1-p)pe^{\xi}/(1-p+pe^{\xi})^2 \leq \frac{1}{4}$ for all ξ, so that a second order Taylor's expansion gives $g(\xi) \leq \xi^2/8$. Combining this with the preceding displays, we obtain, for any $h > 0$,

$$\mathrm{P}\Big(\sum_{i=1}^n (X_i - \mathrm{E}(X_i)) \geq t\Big) \leq \exp(-hnt + h^2 n(b-a)^2).$$

The result follows upon choosing $h = 4t/(b-a)^2$. □

Lemma 7.11 *The neighbourhoods* $\{P: \int \psi\, dP < \int \psi\, dP_0 + c\}$ *for* ψ *ranging over the bounded, continuous functions* $\psi: \mathfrak{X} \to \mathbb{R}$ *and* $c > 0$ *form a subbasis for the weak topology on the set* $\mathfrak{M}(\mathfrak{X})$ *of probability measures on a Polish space* $\mathfrak{X}$*. In other words, every open ball around* P_0 *is the union of finite intersections of neighbourhoods of this type.*

Proof This is general topology applied to the definition of the weak topology as the topology generated by the maps $P \mapsto \int \psi\, dP$, for the given set of ψ. □

Exercises

7.1 Show that the posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is consistent (or strongly consistent, respectively) at θ_0 if and only if $\Pi_n(\cdot|\, X^{(n)}) \to_d \delta_{\theta_0}$ in $P_{\theta_0}^{(n)}$-probability (or almost surely $[P_{\theta_0}^{(\infty)}]$, respectively), as $n \to \infty$.

7.2 Suppose that the posterior distribution $\Pi(\cdot|\, X^{(n)})$ of a probability density is consistent relative to the L_1-distance on the parameter set of densities. Show that the posterior mean density $x \mapsto \int p(x)\, d\Pi_n(p|\, X^{(n)})$ is consistent in L_1, as a point estimator for a density.

8

Tests and metric entropy

This chapter presents results on the construction of exponentially powerful hypothesis tests. Such tests play an important role in consistency and rate theorems for posterior distributions.

8.1 Minimax theorem

Let P be a probability measure and let $\mathcal{Q}$ be a collection of probability measures on a measurable space $(\mathfrak{X}, \mathscr{X})$. The *minimax risk for testing* P versus $\mathcal{Q}$ is defined by

$$\pi(P, \mathcal{Q}) = \inf_{\phi}\Bigl(P\phi + \sup_{Q\in\mathcal{Q}} Q(1-\phi)\Bigr), \tag{8.1}$$

where the infimum is taken over all *tests*, i.e. measurable functions $\phi\colon \mathfrak{X} \to [0,1]$. The problem is to give a manageable bound on this risk, or equivalently on its two components, the probabilities of errors of the first kind $P\phi$ and of the second kind $Q(1-\phi)$. We assume throughout that P and $\mathcal{Q}$ are dominated by a σ-finite measure μ, and denote by p and q the densities of the measures P and Q. Let $\mathrm{conv}(\mathcal{Q})$ denote the *convex hull* of $\mathcal{Q}$: the set of all finite convex combinations $\sum_{i=1}^k \lambda_i Q_i$ of elements $Q_i \in \mathcal{Q}$, where $(\lambda_1, \ldots, \lambda_k) \in \mathbb{S}_k$.

The *Hellinger affinity* of two densities p and q is defined as

$$\rho_{1/2}(p,q) = \int \sqrt{p}\sqrt{q}\,d\mu.$$

It is related to the *Hellinger distance* $h(p,q)$ between p and q, whose square is defined by

$$h^2(p,q) = \int (\sqrt{p} - \sqrt{q})^2\,d\mu = 2 - 2\rho_{1/2}(p,q). \tag{8.2}$$

Proposition 8.1 (Minimax theorem for testing) *For dominated probability measures P and $\mathcal{Q}$*

$$\pi(P,\mathcal{Q}) = 1 - \tfrac{1}{2}\|P - \mathrm{conv}(\mathcal{Q})\|_1 \le \sup_{Q\in\mathrm{conv}(\mathcal{Q})} \rho_{1/2}(p,q).$$

Proof The set of test-functions ϕ can be identified with the nonnegative functions in the unit ball Φ of $L_\infty(\mathfrak{X}, \mathscr{X}, \mu)$, which is dual to $L_1(\mathfrak{X}, \mathscr{X}, \mu)$, since μ is σ-finite. The set Φ is compact and Hausdorff with respect to the weak*-topology, by the Banach-Alaoglu theorem (cf. Theorem 3.15 of Rudin (1973)) and weak-* closure of the set of positive functions. Because the map $(\phi, Q) \mapsto P\phi + Q(1-\phi)$ from $L_\infty(\mathfrak{X}, \mathscr{X}, \mu) \times L_1(\mathfrak{X}, \mathscr{X}, \mu)$ to $\mathbb{R}$ is convex and weak*-continuous in ϕ and linear in Q, the minimax theorem (see Theorem 8.11) gives

$$\inf_{\phi\in\Phi}\sup_{Q\in\mathrm{conv}(\mathcal{Q})}\bigl(P\phi+Q(1-\phi)\bigr)=\sup_{Q\in\mathrm{conv}(\mathcal{Q})}\inf_{\phi\in\Phi}\bigl(P\phi+Q(1-\phi)\bigr).$$

The expression on the l.h.s. is the minimax testing risk $\pi(P,\mathcal{Q})$, as replacing $\mathcal{Q}$ by its convex hull does not change the minimax testing risk.

For fixed p,q the expression $P\phi+Q(1-\phi)=1+\int\phi(p-q)\,d\mu$ is minimized over all test functions by choosing ϕ the minimal possible value 0 if $p-q>0$ and equal to the maximal value 1 if $p-q<0$. In other words, the infimum in the right side is attained for $\phi=1\!\!1\{p<q\}$, and the minimal value is equal to $P(p<q)+Q(p\geq q)=1-\int(p-q)^-\,d\mu$. Because $0=\int(p-q)\,d\mu=\int(p-q)^+\,d\mu-\int(p-q)^-\,d\mu$, the latter can be rewritten as $1-\frac{1}{2}\|p-q\|_1$.

For the inequality we write

$$P(p<q)+Q(p\geq q)=\int_{p<q}p\,d\mu+\int_{p\geq q}q\,d\mu,$$

and bound p in the first integral and q in the second by $\sqrt{p}\sqrt{q}$. □

The proposition shows the importance of the convex hull of $\mathcal{Q}$. Not the separation of $\mathcal{Q}$ from the null hypothesis, but the separation of its convex hull drives the error probabilities.

8.2 Product measures

We shall be interested in tests based on n i.i.d. observations. In other words, we shall apply Proposition 8.1 with the general P and Q replaced by product measures P^n and Q^n. Because the L_1-distance between product measures is difficult to handle, the further bound by the Hellinger affinity is useful. By Fubini's theorem this is multiplicative in product measures:

$$\rho_{1/2}(p_1\times p_2,q_1\times q_2)=\rho_{1/2}(p_1,q_1)\rho_{1/2}(p_2,q_2).$$

When we take the supremum over sets of densities, then this multiplicativity is lost, but the following lemma shows that the Hellinger affinity is still "sub-multiplicative".

For $i=1,\ldots,n$ let P_i and $\mathcal{Q}_i$ be a probability measure and a set of probability measures on an arbitrary measurable space $(\mathfrak{X}_i,\mathscr{X}_i)$, and consider testing the product $\otimes_i P_i$ versus the set $\otimes_i\mathcal{Q}_i$ of products $\otimes_i Q_i$ with Q_i ranging over $\mathcal{Q}_i$. For simplicity write $\rho_{1/2}(P,\mathcal{Q})$ for $\sup_{Q\in\mathcal{Q}}\rho_{1/2}(P,Q)$.

Lemma 8.2 *For any probability measures P_i and convex classes $\mathcal{Q}_i$ of probability measures*

$$\rho_{1/2}\bigl(\otimes_i P_i,\mathrm{conv}(\otimes_i\mathcal{Q}_i)\bigr)\leq\prod_i\rho_{1/2}(P_i,\mathcal{Q}_i).$$

Proof If suffices to give the proof for $n=2$; the general case follows by repetition. Any measure $Q\in\mathrm{conv}(\mathcal{Q}_1\times\mathcal{Q}_2)$ can be represented by a density of the form $q(x,y)=$

$\sum_j \kappa_j q_{1j}(x) q_{2j}(y)$, for nonnegative constants κ_j with $\sum_j \kappa_j = 1$, and q_{ij} densities of measures belong to $\mathcal{Q}_i$. Then $\rho_{1/2}(p_1 \times p_2, q)$ can be written in the form

$$\int p_1(x)^{1/2}\Bigl(\sum_j \kappa_j q_{1j}(x)\Bigr)^{1/2}\Bigl[\int p_2(y)^{1/2}\Bigl(\frac{\sum_j \kappa_j q_{1j}(x) q_{2j}(y)}{\sum_j \kappa_j q_{1j}(x)}\Bigr)^{1/2} d\mu_2(y)\Bigr]\, d\mu_1(x).$$

(If $\sum_j \kappa_j q_{1j}(x) = 0$, the quotient in the inner integral is interpreted as 0.) The inner integral is bounded by $\rho_{1/2}(P_2, \mathcal{Q}_2)$ for every fixed $x \in \mathfrak{X}$, since $\mathcal{Q}_2$ is convex by assumption and the function of y within the brackets is for every fixed x a convex combination of the densities q_{2j} (with weights proportional to $\kappa_j q_{1j}(x)$). After substitution of this upper bound the remaining integral is bounded by $\rho_{1/2}(P_1, \mathcal{Q}_1)$, since $\mathcal{Q}_1$ is convex. □

Combining the preceding lemma with Proposition 8.1, we see that, for every *convex* set $\mathcal{Q}$ of measures:

$$\pi(P^n, \mathcal{Q}^n) \le \rho_{1/2}(P^n, \mathcal{Q}^n) \le \rho_{1/2}(P, \mathcal{Q})^n.$$

Thus any convex set $\mathcal{Q}$ with Hellinger affinity to P smaller than 1 can be tested with exponential error probabilities.

Theorem 8.3 *For any probability measure P and convex set of dominated probability measures $\mathcal{Q}$ with $h(p, q) > \epsilon$ for every $q \in \mathcal{Q}$ and any $n \in \mathbb{N}$, there exists a test ϕ such that*

$$P^n \phi \le e^{-n\epsilon^2/2}, \qquad \sup_{Q \in \mathcal{Q}} Q^n(1 - \phi) \le e^{-n\epsilon^2/2}.$$

Proof By (8.2) we have $\rho_{1/2}(P, \mathcal{Q}) = 1 - \frac{1}{2}h^2(P, \mathcal{Q})$, which is bounded above by $1 - \epsilon^2/2$ by assumption. Combined with the display preceding the theorem we see that $\pi(P^n, \mathcal{Q}^n) \le (1 - \epsilon^2/2)^n \le e^{-n\epsilon^2/2}$, since $1 - x \le e^{-x}$, for every x. □

8.3 Tests and entropy

The alternatives that we need to test are complements of balls and are *not* convex. We handle these by covering them with convex sets, and combining the corresponding tests into a single overall test. The power will then depend on the number of sets needed in a cover.

Definition 8.4 (Covering number) Given a semi-metric d on a set $\mathcal{Q}$ and $\epsilon > 0$ the *covering number* $N(\epsilon, \mathcal{Q}, d)$ is defined as the minimal number of balls of radius ϵ needed to cover $\mathcal{Q}$. The logarithm of the covering number is called (metric) *entropy*.

The covering number increases as ϵ decreases to zero. Except in trivial cases, they increase to infinity. The rate of increase is a measure of the size of $\mathcal{Q}$. Section 8.4 contains examples of covering numbers.

Proposition 8.5 *Let d be a metric whose balls are convex and which is bounded above by the Hellinger distance h. If $N(\epsilon/4, \mathcal{Q}, d) \le N(\epsilon)$ for every $\epsilon > \epsilon_n > 0$ and some nonincreasing function $N\colon (0, \infty) \to (0, \infty)$, then for every $\epsilon > \epsilon_n$ and n there exists a test ϕ such that, for all $j \in \mathbb{N}$,*

$$P^n \phi \le N(\epsilon) \frac{e^{-n\epsilon^2/2}}{1 - e^{-n\epsilon^2/8}}, \qquad \sup_{Q \in \mathcal{Q}: d(P,Q) > j\epsilon} Q^n(1 - \phi) \le e^{-n\epsilon^2 j^2/8}.$$

Proof For a given $j \in \mathbb{N}$, choose a maximal set of points $Q_{j,1}, \ldots, Q_{j,N_j}$ in the set $\mathcal{Q}_j := \{Q \in \mathcal{Q}: j\epsilon < d(P,Q) < 2j\epsilon\}$ such that $d(Q_{j,k}, Q_{j,l}) \geq j\epsilon/2$ for every $k \neq l$. Because every ball in a cover of $\mathcal{Q}_j$ by balls of radius $j\epsilon/4$ then contains at most one $Q_{j,l}$, it follows that $N_j \leq N(j\epsilon/4, \mathcal{Q}_j, d)$. Furthermore, the N_j balls $B_{j,l}$ of radius $j\epsilon/2$ around the $Q_{j,l}$ cover $\mathcal{Q}_j$, as otherwise this set was not maximal. Since $Q_{j,l} \in \mathcal{Q}_j$, the distance of $Q_{j,l}$ to P is at least $j\epsilon$ and hence $h(P, B_{j,l}) \geq d(P, B_{j,l}) > j\epsilon/2$ for every ball $B_{j,l}$. By Theorem 8.3 there exists a test $\phi_{j,l}$ of P versus $B_{j,l}$ with error probabilities bounded above by $e^{-nj^2\epsilon^2/8}$. Let ϕ be the supremum of all the tests $\phi_{j,l}$ obtained in this way, for $j = 1, 2, \ldots,$ and $l = 1, 2, \ldots, N_j$. Then

$$P^n\phi \leq \sum_{j=1}^{\infty}\sum_{l=1}^{N_j} e^{-nj^2\epsilon^2/8} \leq \sum_{j=1}^{\infty} N(j\epsilon/4, \mathcal{Q}_j, d) e^{-nj^2\epsilon^2/8} \leq N(\epsilon)\frac{e^{-n\epsilon^2/8}}{1 - e^{-n\epsilon^2/8}}$$

and, for every $j \in \mathbb{N}$,

$$\sup_{Q \in \cup_{l>j}\mathcal{Q}_l} Q^n(1-\phi) \leq \sup_{l>j} e^{-nl^2\epsilon^2/8} \leq e^{-nj^2\epsilon^2/8},$$

since for every $Q \in \mathcal{Q}_j$ there exists a test $\phi_{j,l}$ with $1 - \phi \leq 1 - \phi_{j,l}$, by construction. □

One may note from the proof that it suffices that $N(\epsilon)$ upper bounds the smaller covering numbers $N(\epsilon/4, \{Q \in \mathcal{Q}: \epsilon < d(P,Q) < 2\epsilon\}, d)$. The logarithms of the latter numbers are called *Le Cam dimension.* For genuinely nonparametric applications these are rarely essentially smaller than the numbers $N(\epsilon/4, \mathcal{Q}, d)$, but for finite-dimensional models they may be.

8.4 Examples of entropy

Lemma 8.6 *For the norm $\|x\|_1 = \sum_i |x_i|$ on the d-dimensional unit simplex $\mathbb{S}_d$, for $0 < \epsilon \leq 1$,*

$$N(\epsilon, \mathbb{S}_d, \|\cdot\|_1) \leq \left(\frac{5}{\epsilon}\right)^{d-1}. \tag{8.3}$$

Lemma 8.7 *For $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$ and $p \geq 1$, for any M and $\epsilon > 0$,*

$$N(\epsilon, \{x \in \mathbb{R}^d: \|x\|_p \leq M\}, \|\cdot\|_p) \leq \left(\frac{3M}{\epsilon}\right)^d. \tag{8.4}$$

The preceding lemma shows, in particular, that the Le Cam dimension of a ball in d-dimensional Euclidean space satisfies

$$\log N(\epsilon, \{x \in \mathbb{R}^d: \|x\|_p \leq k\epsilon\}, \|\cdot\|_p) \leq d\log(3k).$$

The bound is independent of ϵ, for any fixed k. Thus on Euclidean space this quantity behaves essentially as the dimension.

The preceding bounds show that entropy numbers of sets in Euclidean spaces grow logarithmically. For infinite-dimensional spaces the growth is much faster, as is illustrated by the following examples.

Lemma 8.8 *For $\|\boldsymbol{\theta}\|_2 = \bigl(\sum_{i=1}^\infty \theta_i^2\bigr)^{1/2}$ the norm of ℓ_2, for all $\epsilon > 0$,*

$$\log N\Bigl(\epsilon, \{\boldsymbol{\theta} \in \ell_2 : \sum_{i=1}^\infty i^{2q}\theta_i^2 \le B^2\}, \|\cdot\|_2\Bigr) \le \log(4(2e)^{2q})\Bigl(\frac{3B}{\epsilon}\Bigr)^{1/q}. \tag{8.5}$$

The *Hölder norm* of order α of a continuous function $f: \mathfrak{X} \to \mathbb{R}$ on a bounded subset $\mathfrak{X} \subset \mathbb{R}^d$ is defined as

$$\|f\|_\alpha = \max_{k:|k|\le m} \sup_{x\in D} |D^k f(x)| + \max_{k:|k|=m} \sup_{x,y\in D: x\ne y} \frac{|D^k f(x) - D^k f(y)|}{\|x-y\|^{\alpha-m}}. \tag{8.6}$$

Here m is the biggest integer strictly smaller than α, and for a vector $k = (k_1, \dots, k_n)$ of integers, D^k is the partial differentiable operator

$$D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}.$$

Lemma 8.9 *There exists a constant K depending only on d and α such that*

$$\log N\bigl(\epsilon, \{f: \|f\|_\alpha \le M\}, \|\cdot\|_\infty\bigr) \le K \text{meas}(\mathfrak{X}) \Bigl(\frac{M}{\epsilon}\Bigr)^{d/\alpha}.$$

Lemma 8.10 *The collection $\mathfrak{F}$ of monotone functions $f: \mathfrak{X} \to [-M, M]$ on an interval $\mathfrak{X} \subset \mathbb{R}$ satisfies, for $\|\cdot\|_{r,Q}$ the $L_r(Q)$ norm relative to a probability measure Q, any $r \ge 1$ and a constant K that depends on r only,*

$$\log N\bigl(\epsilon, \mathfrak{F}, \|\cdot\|_{r,Q}\bigr) \le K\frac{M}{\epsilon}.$$

8.5 COMPLEMENTS

Theorem 8.11 (Minimax theorem) *Let T be a compact, convex set of a locally convex topological vector space (for instance a normed linear space) and S a convex subset of a linear space. Let $f: T \times S \to \mathbb{R}$ be a function such that*

(i) *$t \mapsto f(t,s)$ is continuous and concave for all $s \in S$;*
(ii) *$s \mapsto f(t,s)$ is convex for all $s \in S$.*

Then

$$\inf_{s\in S} \sup_{t\in T} f(t,s) = \sup_{t\in T} \inf_{s\in S} f(t,s). \tag{8.7}$$

For a proof, see Strasser (1985), pages 239–241.

Exercises

8.1 Prove Lemma 8.7. [Hint: given N points $x_1, \dots, x_N$ in $U = \{x: \|x\| < k\epsilon\}$ with $\|x_i - x_j\| > \epsilon$, the balls $\{x: \|x - x_i\| < \epsilon/2\}$ are disjoint and their union is contained in $\{x: \|x\| < (k + 1/2)\epsilon\}$. Now use a volume argument to bound N.]

8.2 Consider the set $\mathcal{F}$ of functions $f\colon [0,1] \to [0,1]$ such that $\bigl|f(x) - f(y)\bigr| \leq |x - y|$, for every $x, y \in [0,1]$. Show that there exists a constant K such that $\log N\bigl(\epsilon, \mathcal{F}, \|\cdot\|_\infty\bigr) \leq K(1/\epsilon)$, for $\epsilon < 1$. [This is a special case of Lemma 8.9. Give a direct proof. Use balls around piecewise constant (or linear) functions.]

8.3 Suppose d_1 and d_2 are metrics with $d_1 \leq d_2$. Show that $N(\epsilon, \mathcal{Q}, d_1) \leq N(\epsilon, \mathcal{Q}, d_2)$, for every $\epsilon > 0$.

9

Rate of contraction

9.1 Definition

For every $n \in \mathbb{N}$ let $X^{(n)}$ be an observation in a sample space $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)})$ with distribution $P_\theta^{(n)}$ indexed by a parameter θ belonging to a metric space Θ. Given a prior Π on the Borel sets of Θ, let $\Pi_n(\cdot|\, X^{(n)})$ be a version of the posterior distribution.

Definition 9.1 (Posterior rate of contraction) The posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is said to *contract at rate* $\epsilon_n \to 0$ at $\theta_0 \in \Theta$ if $\Pi_n\bigl(\theta\colon d(\theta, \theta_0) > M_n\epsilon_n|\, X^{(n)}\bigr) \to 0$ in $P_{\theta_0}^{(n)}$-probability, for every $M_n \to \infty$ as $n \to \infty$.

A rough interpretation of the rate ϵ_n is that the posterior distribution concentrates on balls of radius "of the order ϵ_n" around θ_0. The somewhat complicated construction using the additional sequence M_n expresses the "of the order" part of this assumption. For "every $M_n \to \infty$" must be read as "whenever $M_n \to \infty$, no matter how slowly". Actually, in most nonparametric applications the fixed sequence $M_n = M$ for a large constant M also works. (In many parametric applications, the posterior distribution tends after scaling to a distribution that is supported on the full space, and the $M_n \to \infty$ is important.)

If ϵ_n is a rate of contraction, then every sequence that tends to zero at a slower rate is also a contraction rate, according to the definition. Saying that contraction rate is *at least* ϵ_n would be appropriate. Naturally we are interested in the fastest contraction rate, but we are typically satisfied with knowing some rate that is valid for every θ_0 in a given class of true parameters.

We may view the rate of contraction as the natural refinement of consistency. Consistency requires that the posterior distribution contracts to within arbitrarily small distance ϵ to the true parameter θ_0; the rate as defined here quantifies "arbitrarily small". Typically contraction rates are much more informative about the quality of a Bayesian procedure than is revealed by mere consistency.

An appropriate summary of the location of the posterior distribution inherits its rate of contraction. The same summary as used in Proposition 7.2 also works for rates. The proof is also very similar.

Proposition 9.2 (Point estimator) *Suppose that the posterior distribution* $\Pi_n(\cdot|\, X^{(n)})$ *contracts at rate* ϵ_n *at* θ_0 *relative to the metric* d *on* Θ*. Then* $\hat\theta_n$ *defined as the center of a (nearly) smallest ball that contains posterior mass at least* $1/2$ *satisfies* $d(\hat\theta_n, \theta_0) = O_P(\epsilon_n)$ *under* $P_{\theta_0}^{(n)}$.

In particular, the posterior distribution cannot contract faster than the best point estimator.

This makes it possible to connect the theory of posterior contraction rates to the theory of "optimal" rates of estimation, which are typically defined by the *minimax criterion.*

9.2 Basic contraction theorem

Let the observations be a random sample $X_1, \ldots, X_n$ from a density p that belongs to a set of densities $\mathcal{P}$, relative to a given σ-finite measure ν. Let Π_n be a prior on $\mathcal{P}$, and let p_0 denote the true density of the observations.

Let d be a distance on $\mathcal{P}$ that is bounded above by the Hellinger distance, and set

$$K(p_0;p) = P_0 \log\frac{p_0}{p}, \qquad V(p_0;p) = P_0\Bigl(\log\frac{p_0}{p}\Bigr)^2. \tag{9.1}$$

The first is the Kullback-Leibler divergence, the second a corresponding second moment.

Theorem 9.3 *The posterior distribution contracts at rate ϵ_n at P_0 for any ϵ_n such that $n\epsilon_n^2 \to \infty$ and such that, for positive constants c_1, c_2 and sets $\mathcal{P}_n \subset \mathcal{P}$,*

$$\log N(\epsilon_n, \mathcal{P}_n, d) \le c_1 n\epsilon_n^2, \tag{9.2}$$

$$\Pi_n\bigl(p\colon K(p_0;p) < \epsilon_n^2, V(p_0;p) < \epsilon_n^2\bigr) \ge e^{-c_2 n\epsilon_n^2}, \tag{9.3}$$

$$\Pi_n(\mathcal{P} - \mathcal{P}_n) \le e^{-(c_2+4)n\epsilon_n^2}. \tag{9.4}$$

Proof For every $\epsilon > 4\epsilon_n$ we have $\log N(\epsilon/4, \mathcal{P}_n, d) \le \log N(\epsilon_n, \mathcal{P}_n, d) \le c_1 n\epsilon_n^2$, by assumption (9.2). Therefore, by Proposition 8.5 applied with $N(\epsilon) = \exp(c_1 n\epsilon_n^2)$ (constant in ϵ) and $\epsilon = M\epsilon_n$ and $j = 1$ in its assertion, where $M \ge 4$ is a large constant to be chosen later, there exist tests ϕ_n with errors

$$P_0^n\phi_n \le e^{c_1 n\epsilon_n^2}\frac{e^{-nM^2\epsilon_n^2/8}}{1 - e^{-nM^2\epsilon_n^2/8}}, \qquad \sup_{p\in\mathcal{P}_n: d(p,p_0)>M\epsilon_n} P^n(1-\phi_n) \le e^{-nM^2\epsilon_n^2/8}.$$

For $M^2/8 > c_1$ the first tends to zero. For A_n the event $\bigl\{\int\prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_n(p) \ge e^{-(2+c_2)n\epsilon_n^2}\bigr\}$ we can bound $\Pi_n\bigl(p\colon d(p,p_0) > M\epsilon_n \,|\, X_1, \ldots, X_n\bigr)$ by

$$\phi_n + 1\!\!1\{A_n^c\} + e^{(2+c_2)n\epsilon_n^2}\int_{d(p,p_0)>M\epsilon_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p)(1-\phi_n).$$

The expected values under P_0^n of the first terms tends to zero. The same is true for the second term, by Lemma 9.4 (below). We split the integral in the third term in parts over $\mathcal{P}_n$ and its complement.

The first is

$$P_0^n\int_{p\in\mathcal{P}_n: d(p,p_0)>M\epsilon_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p) \le \int_{p\in\mathcal{P}_n: d(p,p_0)>M\epsilon_n} P^n(1-\phi_n)\, d\Pi_n(p),$$

which is bounded by $e^{-nM^2\epsilon_n^2/8}$, by the construction of the test. The second is bounded by

$$P_0^n\int_{\mathcal{P}-\mathcal{P}_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p) \le \Pi_n(\mathcal{P}-\mathcal{P}_n).$$

This is bounded above by (9.4). For $M^2/8 > 2 + c_2$ all terms tend to zero. □

The condition $n\epsilon_n^2 \to \infty$ excludes the parametric rate $\epsilon_n = n^{-1/2}$, and merely says that we are considering the nonparametric situation, where slower rates obtain. Besides, the theorem characterizes the rate by three conditions.

The last one (9.4) is trivially satisfied by choosing $\mathcal{P}_n = \mathcal{P}$ for every n. Similar as in the consistency theorems the condition expresses that a subset $\mathcal{P} - \mathcal{P}_n$ of the model $\mathcal{P}_n$ that receives very little prior mass does not play a role in the rate of contraction.

The remaining pair (9.2)-(9.3) of conditions is more structural. For given $\mathcal{P}_n$ and c_1, c_2 each of the two conditions on its own determines a minimal value of ϵ_n (as their left sides decrease and their right sides increase if ϵ_n is replaced by a bigger value). The rate of contraction is the slowest one defined by the two inequalities. Condition (9.3) involves the prior, whereas condition (9.2) does not.

Condition (9.3) gives a lower bound on the amount of mass that the prior puts near the true density p_0. The posterior would not contract to p_0 at all if this mass were zero. "Nearness" to p_0 is measured by the Kullback-Leibler divergence $K(p_0;p)$ and a corresponding variance $V(p_0;p)$. Both quantities should be thought of as "quadratic" discrepancies, and $\{p\colon K(p_0;p) < \epsilon^2, V(p_0;p) < \epsilon^2\}$ as a neighbourhood of "size" ϵ. (For instance, if the likelihood ratios p_0/p are bounded away from zero and infinity, then these neighbourhoods are like Hellinger balls of radius ϵ. See Lemmas 9.8 and 9.9.) For nonparametric rates $\epsilon_n \gg n^{-1/2}$, the exponent $n\epsilon_n^2$ will tend to infinity, and the lower bound on the prior mass in (9.3) will be exponentially small.

Condition (9.2) involves the model $\mathcal{P}_n$, and not the prior. It gives an error bound on the complexity of this model, and may be viewed as bound on the precision of recovery of p_0 for *any* statistical procedure, not only Bayesian. Indeed, for d the Hellinger distance, and under some conditions, the solution to inequality (9.2) can be shown to give the *minimax rate* of estimating p_0 given the model $\mathcal{P}_n$. It goes back to non-Bayesian results by Le Cam (1973, 1986) and Birgé (1983) (also see Yang and Barron (??)). As shown in the proof, technically condition (9.2) ensures existence of tests, as discussed in Chapter 8.

The two conditions (9.2)-(9.3) can be connected, and then send the message that a good prior spreads "uniformly" over the model. Consider placing a maximal set of points $p_1, \ldots, p_N$ in $\mathcal{P}_n$ with $d(p_i, p_j) \geq \epsilon_n$. Maximality implies that the balls of radius ϵ_n around the points cover $\mathcal{P}_n$, whence $N \geq N(\epsilon_n, \mathcal{P}_n, d) \geq e^{c_1 n\epsilon_n^2}$, under (9.2). The balls of radius $\epsilon_n/2$ around the points are disjoint and hence the sum of their prior masses will be less than 1. If the prior mass were evenly distributed over these balls, then each would have no more mass than $e^{-c_1 n\epsilon_n^2}$. This is of the same order as the lower bound in (9.3).

The neighbourhood in (9.3) is not a d-ball, and different constants are involved in the two conditions . However, the argument suggests that (9.3) can only be satisfied for every p_0 in the model if the prior "distributes its mass uniformly, at discretization level ϵ_n". This is a heuristic argument only. Refinements of the theorem show that condition (9.3) is stronger than needed.

Lemma 9.4 *For any probability measure* Π *on* $\mathcal{P}$*, and positive constant* ϵ*, with* P_0^n*-probability at least* $1 - (n\epsilon^2)^{-1}$*,*

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi\big(p\colon K(p_0;p) < \epsilon^2, V(p_0;p) < \epsilon^2\big) e^{-2n\epsilon^2}.$$

Proof The integral becomes smaller by restricting it to the set $B{:}=\{p{:}\,K(p_0;p)<\epsilon_n^2, V(p_0;p)<\epsilon_n^2\}$. By next dividing the two sides of the inequality by $\Pi(B)$, we can rewrite the inequality in terms of the prior Π restricted and renormalized to a probability measure on B. Thus we may without loss generality assume that $\Pi(B)=1$. By Jensen's inequality applied to the logarithm,

$$\log\int\prod_{i=1}^n\frac{p}{p_0}(X_i)\,d\Pi(P)\geq\sum_{i=1}^n\int\log\frac{p}{p_0}(X_i)\,d\Pi(P)=:Z.$$

The right side has mean $-n\int K(p_0;p)\,d\Pi(p)>-n\epsilon^2$ by the definition of B, and variance bounded above by

$$nP_0\Bigl(\int\log\frac{p_0}{p}\,d\Pi(p)\Bigr)^2\leq nP_0\int\Bigl(\log\frac{p_0}{p}\Bigr)^2\,d\Pi(p)\leq n\epsilon^2,$$

by Jensen's inequality, Fubini's theorem, and again the definition of B. It follows that

$$P_0^n\bigl(Z<-2n\epsilon^2\bigr)\leq P_0^n\bigl(Z-\mathrm{E}Z<-n\epsilon^2\bigr)\leq\frac{n\epsilon^2}{(n\epsilon^2)^2},$$

by Chebyshev's inequality. □

9.3 Refinements

[To be skipped at first reading!]

Theorem 9.5 (Almost sure contraction) *If condition (9.3) of Theorem 9.3 is strengthened to*

$$\Pi_n\Bigl(p{:}\,h^2(p,p_0)\Bigl\|\frac{p_0}{p}\Bigr\|_\infty\leq\epsilon_n^2\Bigr)\geq e^{-c_2n\epsilon_n^2},\tag{9.5}$$

and $\sum_{n=1}^\infty e^{-\beta n\epsilon_n^2}<\infty$ *for every* $\beta>0$, *then* $\Pi_n(p{:}\,d(p,p_0)\geq M\epsilon_n|\,X_1,\ldots,X_n)\to 0$, *almost surely* $[P_0^n]$, *for every sufficiently large* M.

Let $B(\epsilon)=\{p{:}\,K(p_0;p)<\epsilon^2, V(p_0;p)<\epsilon^2\}$.

Theorem 9.6 *The posterior distribution contracts at rate* ϵ_n *at* P_0 *for any* $\epsilon_n\geq n^{-1/2}$ *such that, for every sufficiently large* j *and sets* $\mathcal{P}_n\subset\mathcal{P}$,

$$\sup_{\epsilon>\epsilon_n}\log N\Bigl(\frac{\epsilon}{2},\{p\in\mathcal{P}_n{:}\,\epsilon\leq d(p,p_0)\leq 2\epsilon\},d\Bigr)\leq n\epsilon_n^2,,\tag{9.6}$$

$$\frac{\Pi_n(\mathcal{P}-\mathcal{P}_n)}{\Pi_n\bigl(B(\epsilon_n)\bigr)}=o\bigl(e^{-2n\epsilon_n^2}\bigr),\tag{9.7}$$

$$\frac{\Pi_n\bigl(P{:}\,j\epsilon_n<d(P,P_0)\leq 2j\epsilon_n\bigr)}{\Pi_n\bigl(B(\epsilon_n)\bigr)}\leq e^{n\epsilon_n^2j^2/8}.\tag{9.8}$$

Theorem 9.7 *The posterior distribution contracts at rate* ϵ_n *at* P_0 *for any* ϵ_n *with* $n\epsilon_n^2\to$

∞ *that satisifes* (9.3) *such that there exist* $\mathcal{P}_n \subset \mathcal{P}$ *with* $\Pi_n(\mathcal{P}_n^c|\, X_1,\ldots,X_n) \to 0$ *in* P_0^n*-probability and partitions* $\mathcal{P}_n = \cup_{j=-\infty}^{\infty}\mathcal{P}_{n,j}$ *in sets such that*

$$\sum_{j=-\infty}^{\infty}\sqrt{N(\epsilon_n,\mathcal{P}_{n,j},d)}\sqrt{\Pi_n(\mathcal{P}_{n,j})}e^{-n\epsilon_n^2} \to 0.$$

9.4 COMPLEMENTS

The following results show among others that the Kullback-Leibler always dominates the square Hellinger distance, and the converse is true if the likelihood ratios are bounded.

Lemma 9.8 *For any pair of probability densities* p, q,

(i) $\|p-q\|_1 \le h(p,q)\sqrt{4-h^2(p,q)} \le 2h(p,q)$.
(ii) $h^2(p,q) \le \|p-q\|_1$.
(iii) $\|p-q\|_1^2 \le 2K(p;q)$. *(*Kemperman's inequality*)*.
(iv) $h^2(p,q) \le K(p;q)$.
(v) $h(p,q) \le \big\|(\sqrt{p}+\sqrt{q})^{-1}\big\|_\infty \|p-q\|_2$.
(vi) $\|p-q\|_2 \le \big\|\sqrt{p}+\sqrt{q}\big\|_\infty h(p,q)$.
(vii) $\|p-q\|_r \le \|p-q\|_\infty \nu(\mathfrak{X})^{1/r}$, *for* $r \ge 1$.

Lemma 9.9 *For every* $b>0$*, there exists a constant* $\epsilon_b > 0$ *such that for all probability densities* p *and* q *with* $0 < h^2(p,q) < \epsilon_b P(p/q)^b$,

$$K(p,q) \lesssim h^2(p,q)\Big(1 + b^{-1}\log_- h(p,q) + b^{-1}\log_+ P\Big(\frac{p}{q}\Big)^b\Big),$$
$$V(p,q) \lesssim h^2(p,q)\Big(1 + b^{-1}\log_- h(p,q) + b^{-1}\log_+ P\Big(\frac{p}{q}\Big)^b\Big)^2.$$

Consequently, for every pair of probability densities p *and* q,

$$K(p;q) \lesssim 2h^2(p,q)\Big(1+\log\Big\|\frac{p}{q}\Big\|_\infty\Big) \le 2h^2(p,q)\Big\|\frac{p}{q}\Big\|_\infty,$$
$$V(p;q) \lesssim h^2(p,q)\Big(1+\log\Big\|\frac{p}{q}\Big\|_\infty\Big)^2 \le 2h^2(p,q)\Big\|\frac{p}{q}\Big\|_\infty.$$

10

Gaussian process priors

Gaussian processes are widely used in Bayesian nonparametrics as building blocks for prior models for unknown functional parameters. In this chapter we define such processes, study basic properties and give important examples.

10.1 Stochastic process priors

We start with recalling the general definition of a stochastic process.

Definition 10.1 Let T be a set and $(E, \mathcal{E})$ a measurable space. A *stochastic process* indexed by T, taking values in $(E, \mathcal{E})$, is a collection $X = (X_t \colon t \in T)$ of measurable maps X_t from a probability space $(\Omega, \mathscr{U}, \mathrm{Pr})$ to $(E, \mathcal{E})$. The space $(E, \mathcal{E})$ is called the *state space* of the process.

Although in Bayesian nonparametrics applications it is usually purely artificial, we think of the index t as a time parameter, and view the index set T as the set of all possible time points. The state space $(E, \mathcal{E})$ will most often simply be the real line $\mathbb{R}$, endowed with its Borel σ-algebra.

For every fixed $t \in T$ the stochastic process X gives us an E-valued random element X_t on $(\Omega, \mathscr{U}, \mathrm{Pr})$. We can also fix $\omega \in \Omega$ and consider the map $t \mapsto X_t(\omega)$ on T. These maps are called the *trajectories*, or *sample paths* of the process. The sample paths are functions from T to E, i.e. elements of E^T. Hence, we can view the process X as a random element of the function space E^T. Quite often, the sample paths are in fact elements of a nice subset $\Theta \subset E^T$, for instance a space of continuous functions, or functions with a certain degree of smoothness. The process X can then be viewed as a measurable map $X \colon (\Omega, \mathscr{U}, \mathrm{Pr}) \to (\Theta, \mathscr{B})$, where $\mathscr{B}$ is some natural σ-algebra on Θ. In that case the distribution, or *law* of X is the probability measure on $(\Theta, \mathscr{B})$ defined by $B \mapsto \mathrm{Pr}(X \in B)$ for $B \in \mathscr{B}$. If a statistical model is indexed by a function $\theta \in \Theta$, and the likelihood is appropriately measurable, then the law of the process X can be used as a prior distribution.

Definition 10.2 A prior distribution arising from a stochastic process in this manner is called a *stochastic process prior*.

Random measures can be viewed as stochastic process priors, cf. Section 3.3. A stochastic process that is also often used to construct priors is the Brownian motion process, which is defined as follows.

Definition 10.3 The stochastic process $W = (W_t \colon t \geq 0)$ is called a (standard) *Brownian motion*, or *Wiener process*, if

(i) $W_0 = 0$ a.s.,
(ii) $W_t - W_s$ is independent of $(W_u: u \leq s)$ for all $s \leq t$,
(iii) $W_t - W_s$ has a $N(0, t-s)$-distribution for all $s \leq t$,
(iv) almost all sample paths of W are continuous.

Property (i) says that a standard Brownian motion *starts in* 0. A process with property (ii) is called a process with *independent increments*. Property (iii) implies that that the distribution of the increment $W_t - W_s$ only depends on $t - s$. This is called the *stationarity of the increments*. A stochastic process which has property (iv) is called a *continuous process*.

It is not clear from the definition that the Brownian motion actually exists. Observe however that items (i)–(iii) of the definition are equivalent to the requirement that for every $n \in \mathbb{N}$ and all $0 \leq t_1 \leq \cdots \leq t_n$, the vector $(W_{t_1}, \ldots, W_{t_n})$ has an n-dimensional normal distribution with mean 0 and covariance matrix with elements

$$\mathrm{E} W_{t_i} W_{t_j} = t_i \wedge t_j$$

(check!). Proposition 3.7 then implies that there exists a stochastic process W that satisfies properties (i)–(iii) (see Exercise 10.1).

Item (iv) of the definition requires more care. Given a process W that satisfies (i)–(iii), the set $\{\omega: t \mapsto W_t(\omega) \text{ is continuous}\}$ is not necessarily measurable, hence the probability that W has continuous sample paths is not necessarily well defined. However, the continuity criterion of Kolmogorov, Theorem 10.20, implies that the process W that satisfies (i)–(iii) admits a continuous *modification*, i.e. there exists a continuous process $\tilde{W}$ on the same underlying probability space such that for every $t \geq 0$, $W_t = \tilde{W}_t$, almost surely. Note that the process $\tilde{W}$ still satisfies (i)–(iii).

The continuity theorem actually says more, namely that the continuous version is locally Hölder continuous of every order strictly less than $1/2$. Recall that we call a real-valued function f on an interval T *Hölder continuous* of order $\alpha \in (0, 1]$ if there exists a constant $C > 0$ such that $|f(t) - f(s)| \leq C|t - s|^\alpha$ for all $s, t \in T$. A function on an unbounded interval is called *locally Hölder continuous* of order $\alpha > 0$ if the restriction to every bounded interval is Hölder continuous of order α.

Putting everything together, we arrive at the following existence result (see Exercise 10.2).

Theorem 10.4 *Brownian motion exists. Moreover, there exists a version with sample paths that are locally Hölder continuous of order α, for every $\alpha \in (0, 1/2)$.*

It can be proved that the sample paths of Brownian motion are not locally Hölder of order exactly equal to $1/2$. (In particular, they are non-differentiable functions.) However, although strictly speaking it is inaccurate (in the Hölder sense), we say that the sample paths of Brownian motion have "regularity $1/2$".

Figure 10.1 shows an example of a typical Brownian sample path.

10.2 Gaussian processes

Given a real-valued stochastic process X indexed by a set T, we can consider the collection of all distributions of vectors of the the form $(X_{t_1}, \ldots, X_{t_n})$ for $n \in \mathbb{N}$ and $t_1, \ldots, t_n \in T$. These distributions are called the *finite-dimensional distributions* (fdds) of the process. Two

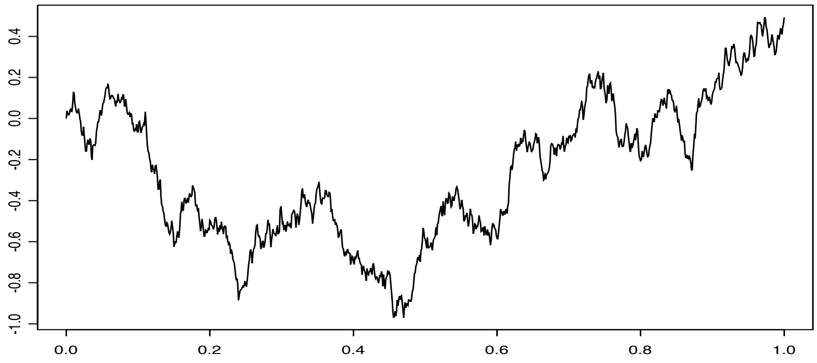

Figure 10.1 A sample path of one-dimensional Brownian motion

processes with the same fdds (not necessarily defined on the same probability space) are called *versions* of each other.

In the preceding section we remarked that the fdds of Brownian motion are all Gaussian. Processes with this property get the obvious name.

Definition 10.5 A real-valued stochastic process is called *Gaussian* if all its fdds are Gaussian.

In other words, a real-valued process $X = (X_t: t \in T)$ is Gaussian if every linear combination $\sum a_i X_{t_i}$, for real numbers $a_1, \ldots, a_n$ and $t_1, \ldots, t_n \in T$, has a Gaussian distribution.

If X is a Gaussian process indexed by the set T, the *mean function* of the process is the function m on T defined by $m(t) = \mathrm{E}X_t$. The *covariance function* of the process is the function r on $T \times T$ defined by $r(s,t) = \mathrm{Cov}(X_s, X_t)$. Note that the functions m and r determine the fdds of the process X, and hence two Gaussian processes with the same mean function and covariance function are versions of each other. We saw that the mean function m and covariance function r of the Brownian motion are given by $m(t) = 0$ and $r(s,t) = s \wedge t$. In general, a Gaussian process X with $\mathrm{E}X_t = 0$ for all $t \in T$ is called *centered.*

The covariance function r of a Gaussian process X is a symmetric, positive definite function. Indeed, for $a_1, \ldots, a_n \in \mathbb{R}$ and $t_1, \ldots, t_n \in T$,

$$\sum\sum a_i a_j r(t_i, t_j) = \mathrm{Var} \sum a_i X_{t_i} \geq 0.$$

Conversely, if T is a set, m is a function on T and r is a symmetric, positive definite function on $T \times T$, then Kolmogorov's extension theorem (Proposition 3.7) implies there exists a Gaussian process X indexed by T, with mean function m and covariance function r.

As mentioned already in connection with the existence of Brownian motion, the question whether the sample paths of a Gaussian process with a given mean and covariance structure have a certain regularity may not be well posed. However, we can give conditions under which a version or modification exists with a certain regularity. A minimal condition on a Gaussian process X indexed by a metric space (T, d) is that it is *mean-square continuous*, which means that for all $s \in T$, $\mathrm{E}(X_t - X_s)^2 \to 0$ as $d(t, s) \to 0$. By Theorem 10.19, a centered, mean-square continuous Gaussian process indexed by a separable metric space has a modification with Borel measurable sample paths.

Stronger assumptions on the second order structure of Gaussian pricess allow to draw stronger conclusions regarding the regularity of sample paths. Suppose for instance that a centered Gaussian process X indexed by a subinterval T of the real line satisfies, for some $K > 0$ and $p \in (0, 1]$, the condition

$$\mathrm{E}(X_s - X_t)^2 \leq K|t - s|^{2p}$$

for all $s, t \in T$. By Gaussianity $X_t - X_s$ has the same distribution as $\big(\mathrm{E}(X_s - X_t)^2\big)^{1/2} Z$ with Z standard normal and hence, for every $a > 0$,

$$\mathrm{E}|X_t - X_s|^a = \mathrm{E}|Z|^a (\mathrm{E}(X_s - X_t)^2)^{a/2} \leq K^{a/2} \mathrm{E}|Z|^a |t - s|^{ap}.$$

By Theorem 10.20 it follows that X admits a modification with sample paths that are a.s. locally Hölder continuous of order α, for every $\alpha \in (0, p)$. (Brownian motion corresponds to the case $p = 1/2$.)

10.3 Examples of Gaussian processes

Brownian motion is a fundamental example of a Gaussian process and we can use it as a building block to construct many more examples.

10.3.1 Wiener integrals

One way to construct new processes from a Brownian W motion is to integrate functions relative to W, that is, to consider integrals of the form $\int f\, dW$. However, since the sample paths of W are very rough, these integrals can not be defined pathwise in the ordinary Lebesgue-Stieltjes sense. The way out is to define them via a Hilbert space isometry. In general this leads to integrals that are not defined pathwise, but only in an L^2-sense.

To have more flexibility we define integrals with respect to a *two-sided* Brownian motion. Let W^1 and W^2 be two independent Brownian motions. Construct a two-sided Brownian motion $W = (W_t\colon t \in \mathbb{R})$, emanating from 0, by setting

$$W_t = \begin{cases} W^1_t & \text{if } t \geq 0, \\ W^2_{-t} & \text{if } t < 0. \end{cases}$$

For real numbers $t_0 < \cdots < t_n$ and $a_1, \ldots, a_n$, consider the simple function $f = \sum a_k 1_{(t_{k-1}, t_k]}$. We define the "integral" of f relative to W in the obvious way by setting

$$\int f\, dW = \sum a_k (W_{t_k} - W_{t_{k-1}}).$$

Using the basic properties of the Brownian motion it is straightforward to verify that for two simple functions f, g, we have

$$\mathrm{E}\Big(\int f\,dW\Big)\Big(\int g\,dW\Big) = \int_{\mathbb{R}} f(x)g(x)\,dx. \tag{10.1}$$

In other words, the linear map $f \mapsto \int f\,dW$ is an isometry from the collection of simple functions in $L^2(\mathbb{R})$ into $L^2(\mathrm{Pr})$. Since the simple functions are dense in $L^2(\mathbb{R})$, the map can be extended to the whole space $L^2(\mathbb{R})$. This defines $\int f\,dW$ for all $f \in L^2(\mathbb{R})$.

Note that by construction the integral is almost surely unique. It is a centered Gaussian random variable and the isometry relation (10.1) holds for all $f, g \in L^2(\mathbb{R})$.

Definition 10.6 We call $\int f\,dW$ the *Wiener integral* of f relative to W. If $f \in L^2(\mathbb{R})$ and $t \geq s$, we write $\int_s^t f(u)\,dW_u$ for $\int 1_{(s,t]} f\,dW$.

Under appropriate conditions, some of the usual calculus rules still hold for the Wiener integral, in particular a version of Fubini's theorem and the integration by parts formula. Recall that we say that $f\colon [s,t] \to \mathbb{R}$ is of bounded variation if

$$\mathrm{var}(f) = \sup \sum |f(t_k) - f(t_{k-1})|$$

is finite, where the supremum is over all finite partitions of $[s,t]$. Note that such a function is necessarily square integrable on $[s,t]$.

Proposition 10.7

(i) (Fubini for Wiener integrals) Let (S, Σ, μ) be a finite measure space and $f \in L^2(\mathrm{Leb} \times \mu)$. Then it almost surely holds that

$$\int \Big(\int f(u,v)\,dW_u\Big)\,\mu(dv) = \int \Big(\int f(u,v)\,\mu(dv)\Big)\,dW_u.$$

(ii) (Integration by parts) If $t \geq s$ and $f\colon [s,t] \to \mathbb{R}$ is of bounded variation, then

$$\int_s^t f(u)\,dW_u = W_t f(t) - W_s f(s) - \int_s^t W_u\,df(u)$$

almost surely.

Proof (i). If f is a simple function of the form $f = \sum a_i 1_{I_i \times E_i}$, for real numbers a_i, intervals I_i and $E_i \in \Sigma$, the statement is trivially true. For a general $f \in L^2(\mathrm{Leb} \times \mu)$, there exists a sequence of simple f_n of the form just described such that $f_n \to f$ in $L^2(\mathrm{Leb} \times \mu)$. Then by Jensen's inequality,

$$\int \Big(\int f_n(u,v)\,\mu(dv) - \int f(u,v)\,\mu(dv)\Big)^2\,du \leq \|f_n - f\|_{L^2}^2 \to 0.$$

Hence, by definition of the Wiener integral,

$$\int \Big(\int f_n(u,v)\,\mu(dv)\Big)\,dW_u \to \int \Big(\int f(u,v)\,\mu(dv)\Big)\,dW_u$$

in $L^2(\mathrm{Pr})$. On the other hand, the convergence $\|f_n - f\|_{L^2}^2 \to 0$ implies that there exists a subsequence n' such that and a set $S' \subset S$ of full μ-measure such that

$$\int (f_n(u,v) - f(u,v))^2\,du \to 0$$

for all $v \in S'$. Again by definition of the Wiener integral it follows that for $v \in S'$,

$$\int f_{n'}(u,v)\,dW_u \to \int f(u,v)\,dW_u$$

in $L^2(\mathrm{Pr})$. First, this implies that there is a further subsequence along the convergence takes place almost surely. Hence, since the left-hand side is a measurable function of v, so is the right-hand side. Second, by Jensen and the ordinary Fubini theorem we have

$$\mathrm{E}\Big(\int\Big(\int f_{n'}(u,v)\,dW_u\Big)\,\mu(dv) - \int\Big(\int f(u,v)\,dW_u\Big)\,\mu(dv)\Big)^2 \to 0.$$

(ii). The function f can be written as the difference of two non-decreasing, cadlag functions on $[s,t]$. Hence it suffices to prove the statement under the assumption that f itself is such a non-decreasing function, so that df is an ordinary Lebesgue-Stieltjes measure. By (i), we then have, a.s.,

$$\begin{aligned}\int_s^t W_u\,df(u) &= \int_s^t\Big(\int_0^s dW_v\Big)\,df(u) + \int_s^t\Big(\int_s^u dW_v\Big)\,df(u)\\ &= (f(t)-f(s))W_s + \int_s^t (f(t)-f(u))\,dW_u.\end{aligned}$$

Rearranging gives the equality we have to prove. □

10.3.2 Riemann-Liouville processes

Brownian motion is a Gaussian processes with "regularity" $1/2$ (cf. Theorem 10.4). A natural way to construct processes with different regularities is to integrate the sample paths of the process one or more times.

Let W be a Brownian motion By Fubini and integration by parts (Proposition 10.7),

$$\int_0^t\int_0^{t_{n-1}}\cdots\int_0^{t_1} W_{t_0}\,dt_0dt_1\cdots t_{n-1} = \frac{1}{n!}\int_0^t (t-s)^n\,dW_s \qquad (10.2)$$

almost surely (see Exercise 10.3). As a process in t, this obviously has regularity $n+1/2$. Now we observe that the right-hand side of (10.2) is not just well defined for $n \in \mathbb{N}$, but for every $n \in \mathbb{R}$ such that $s \mapsto (t-s)^n$ belongs to $L^2[0,t]$, i.e. for every $n > -1/2$. This leads to the definition of the following process, which can be viewed as the $(\alpha - 1/2)$-fold iterated integral of Brownian motion.

Definition 10.8 For $\alpha > 0$ and W a Brownian motion, the process R^α defined by

$$R_t^\alpha = \frac{1}{\Gamma(\alpha+1/2)}\int_0^t (t-s)^{\alpha-1/2}\,dW_s, \qquad t \geq 0,$$

is called a *Riemann-Liouville* process with parameter $\alpha > 0$.

Similar as in (10.2), one has for a suitably integrable function $f\colon [0,T] \to \mathbb{R}$ and $t \in [0,T]$ that

$$\int_0^t\int_0^{t_{n-1}}\cdots\int_0^{t_1} f(t_0)\,dt_0dt_1\cdots t_{n-1} = \frac{1}{(n-1)!}\int_0^t (t-s)^{n-1}f(s)\,ds.$$

Again the right-hand side is well defined for non-integer n as well. For $\alpha > 0$, the operator that maps f to the function $I^\alpha_{0+}f$ defined by

$$(I^\alpha_{0+}f)(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s)\, ds,$$

is called the *Riemann-Liouville operator* of order α. The Riemann-Liouville process derives its name from the fact that for $\alpha > 1/2$, it holds that $R^\alpha = I^{\alpha-1/2}_{0+}W$.

We indeed have that the parameter α describes the regularity of the sample paths of R^α.

Proposition 10.9 *Let $\alpha > 0$. There exists a modification of the process R^α with sample paths that are locally Hölder continuous of the order β, for every $\beta \in (0, \alpha)$.*

Proof We give the proof for $\alpha \in (0,1)$. Consider the processes X and Y defined by

$$X_t = \int ((t-s)_+^{\alpha-1/2} - (-s)_+^{\alpha-1/2})\, dW_s, \quad Y_t = \int_{-\infty}^0 ((t-s)^{\alpha-1/2} - (-s)^{\alpha-1/2})\, dW_s,$$

where $x_+ = \max\{x, 0\}$. Note that $(t-s)_+^{\alpha-1/2} - (-s)_+^{\alpha-1/2} \sim (\alpha - 1/2)t(-s)^{\alpha-3/2}$ for $s \to -\infty$ and $(t-s)_+^{\alpha-1/2} - (-s)_+^{\alpha-1/2} \sim (t-s)_+^{\alpha-1/2}$ for $s \to t$, so that X and Y are well defined. Since, up to constants, R^α is the difference of X and Y, it is enough to show that these two processes have the required regularity.

For the process X we have

$$\begin{aligned} \mathrm{E}(X_t - X_s)^2 &= \int ((t-u)_+^{\alpha-1/2} - (s-u)_+^{\alpha-1/2})^2\, du \\ &= \int ((t-s-u)_+^{\alpha-1/2} - (-u)_+^{\alpha-1/2})^2\, du = \mathrm{E}X_{t-s}^2. \end{aligned}$$

Moreover, it is easy to see that for $t \geq 0$, $\mathrm{E}X_t^2 = t^{2\alpha}\mathrm{E}X_1^2$. Combining these two facts we find that for all s, t,

$$\mathrm{E}(X_t - X_s)^2 = |t-s|^{2\alpha}\mathrm{E}X_1^2.$$

Kolmogorov's continuity criterion thus implies that X has the required modification.

As for Y, we have for $t > s > 0$,

$$\begin{aligned} \mathrm{E}(Y_t - Y_s)^2 &= \int_{-\infty}^0 ((t-u)^{\alpha-1/2} - (s-u)^{\alpha-1/2})^2\, du \\ &= (\alpha - 1/2)^2 \int_{-\infty}^0 \Big(\int_s^t (v-u)^{\alpha-3/2}\, dv \Big)^2\, du \\ &\leq (\alpha - 1/2)^2 (t-s)^2 \int_0^\infty (s+u)^{2\alpha-3}\, du \\ &= (\alpha - 1/2)^2 (t-s)^2 s^{2\alpha-2} \int_0^\infty (1+u)^{2\alpha-3}\, du. \end{aligned}$$

It follows that there exists a constant $C > 0$ such that for $t > s > 0$ such that $|t-s| \leq s$, we have $\mathrm{E}(Y_t - Y_s)^2 \leq C|t-s|^{2\alpha}$. Since it also holds that $\mathrm{E}Y_t^2 = t^{2\alpha}\mathrm{E}Y_1^2$ for all $t \geq 0$, we see that for every $T > 0$ there exists a constant $C_T > 0$ such that $\mathrm{E}(Y_t - Y_s)^2 \leq C_T|t-s|^{2\alpha}$

for all $s, t \in [0, T]$. We can then apply Theorem 10.20 again and conclude that the process Y has a modification with the required regularity as well. □

The process X appearing in the proof of Proposition 10.9 is also a well known process, it is the so-called *fractional Brownian motion* with *Hurst index* α. The process Y is in fact more regular than the process X. One can show that there exists a version of Y which is differentiable on $(0, \infty)$ (see Exercise 10.4).

For completeness we remark that the case $\alpha \geq 1$ in the proof of the preceding proposition can be dealt with by using the properties of the Riemann-Liouville operators. Consider the case $\alpha = 1$ for instance. If W is a Brownian motion, then $I_{0+}^{1/2}W$ is a Riemann-Liouville process with parameter 1. It can be proved that for all $p, q > 0$ such that $p+q \notin \mathbb{N}$, we have $I_{0+}^{p}(C^q[0,1]) \subset C^{p+q}[0,1]$. Since there exists a version of W taking values in $C^\beta[0,1]$ for every $\beta < 1/2$, this implies that there exists a version of R^1 taking values in $C^\beta[0,1]$ for every $\beta < 1$. Larger α's can be dealt with similarly.

10.3.3 Stationary Gaussian processes

In addition to (multiply) integrated Brownian motions, so-called stationary Gaussian processes are widely used in the construction of priors on functions. We begin with an example.

Example 10.10 (Ornstein-Uhlenbeck process) Let W be a two-sided Brownian motion and $\theta, \sigma > 0$. Define a new process X by setting

$$X_t = \sigma \int_{-\infty}^{t} e^{-\theta(t-s)}\, dW_s, \quad t \in \mathbb{R}.$$

The process X is called the *Ornstein-Uhlenbeck process* (OU process) with parameters $\theta, \sigma > 0$. By the isometry property of the Wiener integral we have

$$\mathrm{E}X_sX_t = \sigma^2 \int_{-\infty}^{s\wedge t} e^{-\theta(s-u)}e^{-\theta(t-u)}\, du = \frac{\sigma^2}{2\theta}e^{-\theta|t-s|}.$$

In particular, the OU process is *stationary*. Using Kolmogorov's continuity criterion it can be seen that the process admits a version that is locally Hölder continuous of every order strictly less than $1/2$ (see Exercise 10.6).

Using the *Fourier transform* the covariance function of the OU process can be written in a different way. For $f \in L^2(\mathbb{R})$ the Fourier transform $\hat{f}$ is defined as

$$\hat{f}(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t)e^{i\lambda t}\, dt, \quad \lambda \in \mathbb{R}.$$

If $f \in L^1(\mathbb{R})$ this integral exists in the ordinary sense, if not, its convergence has to be understood in L^2-sense. The *Parseval identity* for the Fourier transform asserts that the transform is an isometry on $L^2(\mathbb{R})$, that is, for $f, g \in L^2(\mathbb{R})$,

$$\langle f, g\rangle_2 = \langle \hat{f}, \hat{g}\rangle_2,$$

where $\langle f, g\rangle_2 = \int f(t)\bar{g}(t)\, dt$ is the usual inner product on $L^2(\mathbb{R})$.

Example 10.11 (Ornstein-Uhlenbeck process, continued) We have

$$\frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\theta(t-u)} e^{i\lambda u}\, du = \frac{1}{\sqrt{2\pi}} \frac{\sigma e^{i\lambda t}}{\theta + i\lambda}.$$

Hence, by Parseval, we have the relation

$$\mathrm{E}X_s X_t = \int e^{i\lambda(t-s)} \frac{\sigma^2}{2\pi(\theta^2 + \lambda^2)}\, d\lambda$$

for the OU process. We say that the OU process has *spectral density* $\lambda \mapsto \sigma^2/((2\pi)(\theta^2 + \lambda^2))$.

The notions encountered in the example of the OU process extend to more general processes and *fields*, i.e. stochastic processes indexed by higher-dimensional spaces.

Definition 10.12 Let $T \subset \mathbb{R}^d$. A centered process $X = (X_t : t \in T)$ with finite second moments is called (wide sense) *stationary* is its covariance function satisfies $\mathrm{E}X_s X_t = r(t-s)$ for some function $r: T \to \mathbb{R}$.

There is a one-to-one correspondence between mean-square continuous, stationary Gaussian processes indexed by $\mathbb{R}^d$ and the class of finite Borel measures μ on $\mathbb{R}^d$ that are *symmetric*, in the sense that $\int f(\lambda)\, \mu(d\lambda) = \int f(-\lambda)\, \mu(d\lambda)$ for all bounded measurable functions f.

Theorem 10.13 *The process X is a mean-square continuous stationary process indexed by $\mathbb{R}^d$ if and only if there exists a finite, symmetric Borel measure μ on $\mathbb{R}^d$ such that*

$$\mathrm{E}X_s X_t = \int e^{i\langle \lambda, t-s \rangle}\, \mu(d\lambda) \tag{10.3}$$

for all $s, t \in \mathbb{R}^d$.

Proof If the covariance function of a process X is given by (10.3), then clearly X is stationary. Moreover, we have

$$\mathrm{E}(X_t - X_s)^2 = \int |e^{i\langle \lambda, t \rangle} - e^{i\langle \lambda, s \rangle}|^2\, \mu(d\lambda) \to 0$$

as $t \to s$, by dominated convergence for instance. Hence, the process is mean-square continuous.

On the other hand, suppose that X is mean-square continuous and stationary. There exists a function $r: \mathbb{R}^d \to \mathbb{R}$ such that $\mathrm{E}X_s X_t = r(t-s)$ for all $s, t \in \mathbb{R}^d$. By Cauchy-Schwarz we have

$$|r(t) - r(s)|^2 = |\mathrm{E}(X_t - X_s)X_0|^2 \le \mathrm{E}(X_t - X_s)^2 \mathrm{E}X_0^2.$$

Hence, the fact that X is mean-square continuous implies that r is a continuous function. It is also a positive-definite function, in the sense that for $a_1, \ldots, a_n \in \mathbb{R}$ and $t_1, \ldots, t_n \in \mathbb{R}^d$,

$$\sum \sum a_i a_j r(t_i - t_j) = \mathrm{Var} \sum a_i X_{t_i} \ge 0.$$

The conclusion of the theorem thus follows from Bochner's theorem (Theorem 10.21). □

Since a finite measure is completely determined by its characteristic function, the measure μ in (10.3) is necessarily unique.

Definition 10.14 The measure μ in (10.3) is called the *spectral measure* of the process X. If it admits a Lebesgue density, this is called the *spectral density*.

The spectral representation (10.3) shows that we have a linear isometry between the closure of the linear span of the random variables $\{X_t : t \in \mathbb{R}\}$ in $L^2(\mathrm{Pr})$ and the closure of the linear span of the functions $\{\lambda \mapsto e_t(\lambda) = \exp(i\lambda t) : t \in \mathbb{R}\}$ in $L^2(\mu)$. The isometry is completely determined by linearity and the association $X_t \leftrightarrow e_t$. We call this isometry the *spectral isometry*. It can often be used to translate probabilistic problems concerning the process X into analytic problems regarding the functions e_t in $L^2(\mu)$.

Any symmetric, finite Borel measure μ on $\mathbb{R}^d$ is the spectral measure of a stationary Gaussian process. Indeed, given μ, the map

$$(s, t) \mapsto \int e^{i\langle \lambda, t-s\rangle}\, \mu(d\lambda)$$

is a symmetric, positive definite function. Hence, by Kolmogorov's extension theorem, there exists a centered Gaussian process X which has this function as its covariance function.

The regularity of a stationary Gaussian process X is determined by the tails of its spectral measure μ. Intuitively, heavier tails means "more high frequencies", which implies that the sample paths of the process are less regular. Suppose for simplicity that $d = 1$. It is not difficult to see that if μ has a finite second moment, then the map $t \mapsto e_t$ is differentiable in $L^2(\mu)$. Indeed, we have that $\lambda \mapsto i\lambda e_t(\lambda)$ belongs to $L^2(\mu)$ in this case and for $t \in \mathbb{R}$,

$$\int \Big| \frac{e_{t+h}(\lambda) - e_t(\lambda)}{h} - i\lambda e_t(\lambda) \Big|^2 \mu(d\lambda) \to 0$$

as $h \to \infty$, by dominated convergence (check!). In view of the spectral isometry this means there exist random variables X'_t for every t, such that

$$\frac{X_{t+h} - X_t}{h} \to X'_t$$

in $L^2(\mathrm{Pr})$ as $h \to 0$. In other words, the process X is differentiable in mean-square sense, with derivative X'. Note that the derivative is again a stationary process, with spectral measure $|\lambda|^2\, \mu(d\lambda)$.

We give two examples of stationary Gaussian processes often used in Bayesian nonparametrics and that can be defined through their spectral measures.

Example 10.15 (Matérn process) For $d \in \mathbb{N}$ and $\alpha > 0$, the *Matérn process* on $\mathbb{R}^d$ with parameter α is defined as the centered stationary process with spectral density

$$\lambda \mapsto \frac{1}{\big(1 + \|\lambda\|^2\big)^{\alpha + d/2}}.$$

As before, the parameter α describes the regularity of the process, we illustrate this in the case $d = 1$. Let k be the smallest integer strictly smaller than α. Then the spectral measure of the Matérn process X has a finite moment of order $2k$, hence it is k times differentiable in mean-square sense and its kth derivative $X^{(k)}$ is a stationary process with spectral density

$$\lambda \mapsto \frac{\lambda^{2k}}{\big(1+\lambda^2\big)^{\alpha+1/2}}.$$

By the spectral representation of $X^{(k)}$,

$$\begin{aligned}\mathrm{E}(X_t^{(k)}-X_s^{(k)})^2 &= \int |e^{i\lambda(t-s)}-1|^2 \frac{\lambda^{2k}}{\big(1+\lambda^2\big)^{\alpha+1/2}}\,d\lambda\\ &= |t-s|^{2(\alpha-k)}\int |e^{i\lambda}-1|^2 \frac{\lambda^{2k}}{\big((t-s)^2+\lambda^2\big)^{\alpha+1/2}}\,d\lambda\\ &\le |t-s|^{2(\alpha-k)}\int |e^{i\lambda}-1|^2|\lambda|^{-1-2(\alpha-k)}\,d\lambda.\end{aligned}$$

If α is not an integer, then $\alpha-k\in(0,1)$ and hence the last integral is a finite constant (check!). Then by Kolmogorov's continuity criterion, $X^{(k)}$ admits a version that is locally Hölder of every order less than $\alpha-k$. Hence, by integration, the original process X admits a version that is locally Hölder of every order strictly less than α. If α is an integer (then $k=\alpha-1$) the preceding upper bound remains true with α replaced by $\alpha-\epsilon$ for small enough $\epsilon>0$. By the same reasoning we conclude that X then has a version that is locally Hölder of every order less than $\alpha-\epsilon$. Since $\epsilon>0$ can be chosen arbitrarily small, we reach the same conclusion.

Note that the Ornstein-Uhlenbeck process is a special instance of the Matérn process, corresponding to the choices $d=1$, $\alpha=1/2$.

Example 10.16 (Squared exponential process) The *squared exponential process* is the zero-mean Gaussian process with covariance function

$$\mathrm{E}W_sW_t = e^{-\|s-t\|^2}, \qquad s,t\in\mathbb{R}^d.$$

Like the Matérn process the squared exponential process is stationary. Its spectral measure is easily found using the well-known fact that the Fourier transform of the Gaussian density is Gaussian again. Specifically, we have that the spectral density of the squared exponential process is given by

$$\lambda\mapsto\frac{1}{2^d\pi^{d/2}}e^{-\|\lambda\|^2/4}.$$

This density has finite moments of every order, hence the sample paths of the square exponential process are infinitely often differentiable (in mean-square sense). Figure 10.16 shows examples of sample paths of the process, which indeed are very smooth.

10.4 Illustration: Gaussian process regression

Suppose we have observations $Y=(Y_1,\ldots,Y_n)$ satisfying a regression relation

$$Y_i=\theta(t_i)+\epsilon_i,$$

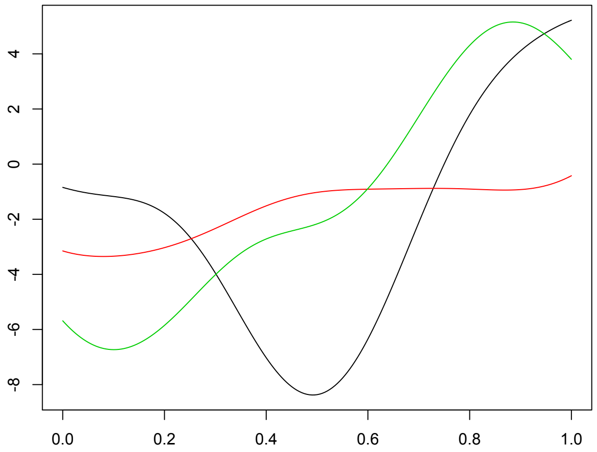

Figure 10.2 Three realizations of the squared exponential process.

where the t_i are fixed, known elements of $[0,1]$, the ϵ_i are independent standard normal variables, and the unknown regression function θ: $[0,1] \to \mathbb{R}$ is the object of interest. We wish to make Bayesian inference about the function θ, which means we have to choose a prior distribution. We take the law Π of an integrated Brownian motion, which we can view as a measure on the space $C[0,1]$ of continuous functions on $[0,1]$.

Under the prior the vector $\theta = (\theta(t_1),\ldots,\theta(t_n))$ (slight abuse of notation!) has a Gaussian distribution with a density equal to a multiple of $\theta \mapsto \exp(-(1/2)\theta^T\Sigma^{-1}\theta)$, for Σ an invertible covariance matrix. The likelihood for this model is given by a multiple of $\exp(-\frac{1}{2}\sum_{i=1}^n (Y_i - \theta(t_i))^2)$. It then follows from Bayes' formula that the posterior density of v is given by a multiple of $\theta \mapsto e^{-\frac{1}{2}\|Y-\theta\|^2} e^{-\frac{1}{2}\theta^T\Sigma^{-1}\theta}$. Clearly this is again, up to a multiple, a Gaussian density. We conclude that the posterior distribution of $(\theta(t_1),\ldots,\theta(t_n))$ is Gaussian process.

Similarly we can show that for *any* sequence $s_1,\ldots,s_m \in [0,1]$, the posterior distributions of $(\theta(s_1),\ldots,\theta(s_m))$ is Gaussian. We conclude that for this Gaussian regression model with known error variances, Gaussian process priors for the regression function are *conjugate* in the sense that the posterior distribution of the regression function is Gaussian as well.

By completing the square, the preceding shows in particular that the posterior distribution of the vector $(\theta(t_1),\ldots,\theta(t_n))$ is Gaussian, with mean vector $(I+\Sigma^{-1})^{-1}Y$ and covariance matrix $(I+\Sigma^{-1})^{-1}$, for Σ the prior covariance matrix of the vector (check!, see Exercise 10.8). For the integrated Brownian motion X we have that for $s \leq t$,

$$\mathrm{E}X_sX_t = \int_0^s \int_0^t (u \wedge v)\, du dv = \frac{1}{2}s^2t - \frac{1}{6}t^3.$$

This gives explicit expression for the matrix Σ and hence the posterior can be computed explicitly.

In Figure 10.4 we show a simulation example. We simulated $n = 200$ observations, with $t_i = i/n$. The black curve depicts the true regression function used in the simulations, the black dots are the simulated noisy data points. In the left panel, 10 draws from the prior are shown. In the right panel, 10 draws from the corresponding posterior are plotted.

In more realistic situations the errors ϵ_i have an unknown variance σ. This additional parameter then has to be endowed with a prior distribution as well. This typically means that the resulting posterior for θ (and for σ) can not be calculated explicitly anymore. There are however numerical methods available that allow us to generate (approximate) draws from the posterior in that case.

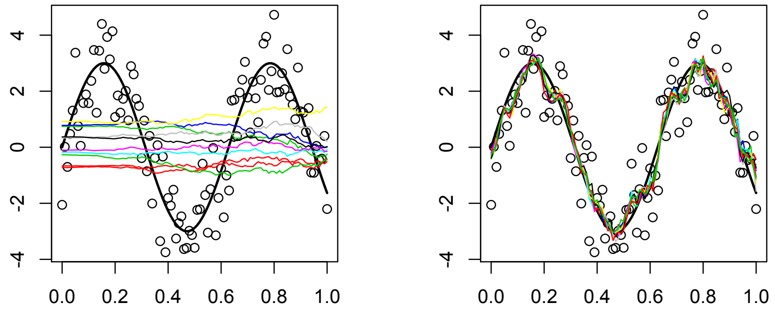

Figure 10.3 Left: 10 draws from the integrated Brownian motion prior (gray), the true regression function and the data. Right: 10 draws from the posterior (gray) and the true regression function and the data.

10.5 COMPLEMENTS

10.5.1 Regular versions of stochastic processes

A minimal regularity property of a stochastic process is *separability*. Roughly speaking, the behaviour of a separable process over the whole index set is determined by its behaviour on a countable subset.

Definition 10.17 Let $(X_t : t \in T)$ be a process indexed by a topological space T, with state space $(E, \mathcal{E})$, where E is a topological space and $\mathcal{E}$ is its Borel σ-field. The process is called *separable* if there exists an event N with $\Pr(N) = 0$ and a countable set $S \subset T$ such that for all open $U \subset T$ and closed $F \subset E$, the subsets $\bigcap_{t\in U}\{X_t \in F\}$ and $\bigcap_{t\in S\cap U}\{X_t \in F\}$ of the underlying outcome space differ by at most a subset of N. Any countable set S with the stated property is called a *separability set*.

The definition immediately implies that for a separable process X indexed by T and defined on a complete probability space, the set $\bigcap_{t\in U}\{X_t \in F\}$ is a measurable event for every open $U \subset T$ and closed $F \subset E$. If the process is real-valued we have in particular that for every $b \in \mathbb{R}$,

$$\Big\{\sup_{t\in T} X_t \le b\Big\} = \bigcap_{t\in T}\{X_t \le b\}$$

is measurable, and hence $\sup_{t\in T} X_t$ is a well-defined random variable. Moreover, it is a.s. equal to $\sup_{t\in T\cap S} X_t$. Similarly, the variables $\inf X_t$, $\sup |X_t|$ and $\inf |X_t|$ are measurable as well.

Another useful consequence of separability is that to show that a real-valued separable process X vanishes identically with probability one, it suffices to show that $X_t = 0$ a.s. for every t in a separability set S. Indeed, suppose that $X_t = 0$ a.s., for all $t \in S$. Then $\bigcap_{t\in S}\{X_t = 0\}$ has probability one. By separability the event $\bigcap_{t\in T}\{X_t = 0\}$ is measurable and has probability one as well, i.e. $\Pr(X_t = 0 \text{ for all } t \in T) = 1$.

In addition to separability it is useful to know whether a stochastic process X is measurable as function of the pair (ω, t). By Fubini's theorem this implies for instance that the sample paths of the process are measurable functions.

Definition 10.18 Let X be a real-valued process indexed by a metric space (T, d), defined on $(\Omega, \mathscr{U}, \Pr)$. Let $\mathbb{B}(T)$ be the Borel σ-field of T. The process X is called *(Borel) measurable*, if the map from $(\Omega \times T, \mathscr{U} \times \mathbb{B}(T))$ to $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ given by $(\omega, t) \mapsto X_t(\omega)$ is measurable.

A process X indexed by a metric space (T, d) is called *continuous in probability* if for all $s \in T$, $X_t \to X_s$ in probability if $d(t, s) \to 0$. The following theorem says that a process admits a measurable and separable modification if it is continuous in probability. Note that this property only depends on the fdds of the process X.

Theorem 10.19 *Let X be a real-valued process indexed by a separable metric space (T, d). If the process is continuous in probability, it admits a Borel measurable, separable version, which may take the value ∞. Any countable, dense subset of (T, d) is a separability set.*

Theorem 10.20 (Kolmogorov's continuity criterion) *Let X be a real-valued process indexed by a compact interval $T \subset \mathbb{R}$. Suppose there exist constants $p, q, C > 0$ such that*

$$\mathrm{E}|X_t - X_s|^p \le C|t - s|^{1+q} \tag{10.4}$$

for all $s, t \in T$. Then X admits a continuous modification with sample paths that are almost surely Hölder continuous of order α for every $\alpha < q/p$.

10.5.2 Bochner's theorem

A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *positive definite* if for all $a_1, \dots, a_n \in \mathbb{R}$ and $t_1, \dots, t_n \in \mathbb{R}^d$,

$$\sum\sum a_i a_j f(t_i - t_j) \ge 0.$$

Bochner's theorem asserts that among the continuous functions on $\mathbb{R}^d$, the positive definite ones are precisely the Fourier transforms of finite measures.

Theorem 10.21 (Bochner) *A continuous function f on $\mathbb{R}^d$ is positive definite if and only if it is given by*

$$f(t) = \int e^{i\langle\lambda,t\rangle}\,\mu(d\lambda), \quad t \in \mathbb{R}^d,$$

for a finite Borel measure μ.

Exercises

10.1 Prove that there exists a stochastic process W that satisfies conditions (i)–(iii) of Definition 10.3.

10.2 Give the details of the proof of Theorem 10.4.

10.3 Verify (10.2).

10.4 Show that there exists a version of the process Y appearing in the proof of Proposition 10.9 which is differentiable on $(0,\infty)$.

10.5 Show that the Riemann-Liouville process R^α with parameter $\alpha > 0$ is *self-similar*: for $c > 0$, the process $(c^{-\alpha}R_{ct}: t \geq 0)$ is again a Riemann-Liouville process with parameter α.

10.6 Prove that the Ornstein-Uhlenbeck process admits a version that is locally Hölder continuous of every order strictly less than $1/2$.

10.7 Show that the Ornstein-Uhlenbeck process satisfies the integral equation

$$X_t - X_0 = -\theta\int_0^t X_s\,ds + \sigma W_t, \quad t \geq 0,$$

almost surely.

10.8 Verify the posterior computations in Section 10.4.

11

Gaussian random elements in Banach spaces

A continuous process $X = (X_t : t \in [0,1]^d)$ defined on $(\Omega, \mathscr{U}, \Pr)$ can also be viewed as a map from Ω to $C[0,1]^d$, where the latter is the space of continuous functions on $[0,1]^d$. Slightly abusing notation this map is denoted by X as well, i.e. $X : \Omega \to C[0,1]^d$. The space $C[0,1]^d$ becomes a separable Banach space when endowed with the uniform norm $\|f\|_\infty = \sup_{t \in [0,1]^d} |f(t)|$. We denote its Borel σ-field, i.e. the σ-field generated by the open sets, by $\mathscr{B}(C[0,1]^d)$.

Proposition 11.1 *For the continuous process X, the map $X : (\Omega, \mathscr{U}) \to (C[0,1]^d, \mathscr{B}(C[0,1]^d))$ is measurable.*

Proof By translation invariance of the topology it suffices to show that for all $a > 0$, it holds that $\{\omega : \|X(\omega)\|_\infty < a\} \in \mathscr{U}$. For a countable, dense subset $D \subset [0,1]^d$, continuity implies that

$$\{\omega : \|X(\omega)\|_\infty < a\} = \bigcap_{t \in D} \{\omega : |X_t(\omega)| < a\}.$$

Every set appearing on the right is in $\mathscr{U}$, since every X_t is a random variable. Since the intersection is countable, it follows that the left-hand side belongs to $\mathscr{U}$ as well. □

The proposition asserts that a continuous process X indexed by $[0,1]^d$ can be viewed as a $(C[0,1]^d, \mathscr{B}(C[0,1]^d))$-valued random element. More generally, it can be shown that a process X taking values in a separable Banach space can be viewed as a (measurable) random element in that space if every bounded linear functional of X is a random variable. In this chapter we will restrict our attention to the case of continuous processes, but the concepts and results we treat can all be extended to the more general Banach space setting.

11.1 Reproducing kernel Hilbert space

Let $X = (X_t : t \in [0,1]^d)$ be a centered, continuous Gaussian process defined on $(\Omega, \mathscr{U}, \Pr)$. We first define the associated space of linear functionals of the process. This is, by definition, the closure in $L^2(\Pr)$ of the collection of linear combinations of the form $\sum a_i X_{t_i}$, for $n \in \mathbb{N}$, $t_1, \ldots, t_n \in [0,1]^d$ and $a_1, \ldots, a_n \in \mathbb{R}$. This space, which we denote by $\mathscr{L}$, is called the *first chaos* of the process X. Note that every element of the first chaos is a centered Gaussian random variable.

For $L \in \mathscr{L}$ we define the function h_L on $[0,1]^d$ by setting $h_L(t) = \mathrm{E} X_t L$. We define the *reproducing kernel Hilbert space (RKHS)* $\mathbb{H}$ associated to the process X by $\mathbb{H} = \{h_L : L \in$

$\mathscr{L}\}$. For h_L in $\mathbb{H}$, the *RKHS-norm* is defined by $\|h_L\|^2_{\mathbb{H}} = \mathrm{E}L^2$. Note that if $h_{L_1} = h_{L_2}$, then $L_1 - L_2$ is orthogonal to every X_t in $L^2(\Pr)$. But then $L_1 - L_2$ is orthogonal to the whole space $\mathscr{L}$, hence $L_1 = L_2$, almost surely. This shows that the RKHS-norm is well defined and that the map $L \mapsto h_L$ defines a Hilbert space isometry between the first chaos $\mathscr{L}$ and the RKHS $\mathbb{H}$. In particular, we see that the RKHS is a separable Hilbert space (Exercise 11.3).

Denote the covariance function of the process X by K, so that $K(s,t) = \mathrm{E}X_sX_t = h_{X_t}(s)$ for $s,t \in [0,1]^d$. Then for $h = h_L \in \mathbb{H}$, we have

$$\langle h, K(\cdot,t)\rangle_{\mathbb{H}} = \langle h_L, h_{X_t}\rangle_{\mathbb{H}} = \mathrm{E}LX_t = h(t), \tag{11.1}$$

for every $t \in [0,1]^d$. The fact that functions in $\mathbb{H}$ can be evaluated at the point t by taking the inner product with $K(\cdot,t)$ is called the *reproducing property*. In this context the covariance function K is also called the *reproducing kernel*.

Since the process X is continuous, it is also mean-square continuous (Exercise 11.2). By the reproducing property and Cauchy-Schwarz it follows that for $h \in \mathbb{H}$ and $t_n \to t$ in $[0,1]^d$ we have

$$|h(t_n) - h(t)|^2 = |\langle h, K(\cdot,t_n) - K(\cdot,t)\rangle_{\mathbb{H}}|^2 = \|h\|^2_{\mathbb{H}}\mathrm{E}(X_{t_n} - X_t)^2 \to 0.$$

Hence every function in the RKHS is continuous, i.e. $\mathbb{H} \subset C[0,1]^d$. Moreover, similar as in the preceding display, we have

$$\|h\|^2_\infty \leq \sigma^2(X)\|h\|^2_{\mathbb{H}}, \tag{11.2}$$

where $\sigma^2(X) = \sup_{t\in[0,1]^d} \mathrm{E}X_t^2$. In other words, the norm of the inclusion map $i\colon \mathbb{H} \to C[0,1]^d$ is bounded by $\sigma(X)$.

The unit ball in $\mathbb{H}$ is denoted by $\mathbb{H}_1 = \{h \in \mathbb{H}\colon \|h\|_{\mathbb{H}} \leq 1\}$. This space is always precompact in $C[0,1]^d$, i.e. it has compact closure.

Theorem 11.2 *The RKHS unit ball $\mathbb{H}_1$ is precompact in $C[0,1]^d$.*

Proof By (11.2) the RKHS unit ball $\mathbb{H}_1$ is uniformly bounded. The function $(s,t) \mapsto \mathrm{E}(X_t - X_s)^2$ is continuous, hence uniformly continuous on the compact set $[0,1]^d \times [0,1]^d$. Since the function vanishes on the diagonal, it follows that for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $\|s-t\| < \delta$, then $\mathrm{E}(X_t - X_s)^2 < \epsilon$. Hence for $h \in \mathbb{H}$, the reproducing property and Cauchy-Schwarz imply that if $\|t-s\| < \delta$, then

$$|h(s) - h(t)|^2 \leq \|h\|^2_{\mathbb{H}}\mathrm{E}(X_t - X_s)^2 < \epsilon.$$

In other words, the unit ball $\mathbb{H}_1$ of the RKHS is uniformly equicontinuous as well. The assertion of the theorem thus follows from the Arzelà-Ascoli theorem (Theorem 11.11). □

11.2 Absolute continuity

The law, or distribution P^X of the centered, continuous Gaussian process X indexed by $[0,1]^d$ is the probability measure on the Borel sets of $C[0,1]^d$ defined by $P^X(B) = \Pr(X \in B)$ for $B \in \mathscr{B}(C[0,1]^d)$. A function $h \in \mathbb{H}$ is continuous, so the shifted process $X + h$ induces a distribution P^{X+h} on $C[0,1]^d$ as well. The following theorem asserts that these distributions are equivalent probability measures. The Radon-Nikodym derivate

dP^{X+h}/dP^X, which is a measurable function on $C[0,1]^d$ can be expressed in terms of the Hilbert space isometry between the RKHS $\mathbb{H}$ and the first chaos $\mathscr{L} \subset L^2(\Omega, \mathscr{U}, \Pr)$. We denote this map by U, so $U: \mathbb{H} \to L^2(\Omega, \mathscr{U}, \Pr)$ is defined by $Uh_L = L$.

Theorem 11.3 (Cameron-Martin) *If $h \in \mathbb{H}$ then P^X and P^{X+h} are equivalent Borel measures on $C[0,1]^d$ and*

$$\frac{dP^{X+h}}{dP^X}(X) = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}$$

almost surely.

Sketch of proof The process X can be written as $X = \sum Z_i h_i$, for $Z_i = Uh_i$ independent, standard normal variables and the h_i an orthonormal basis of the RKHS $\mathbb{H}$. The convergence of the series takes place in $C[0,1]^d$ almost surely. The function $h \in \mathbb{H}$ admits an expansion $h = \sum c_i h_i$ for some $c \in \ell^2$. The series converges in $\mathbb{H}$, but then by (11.2) it converges in $C[0,1]^d$ as well. We can thus write $X + h = \sum (Z_i + c_i) h_i$, convergence taking place in $C[0,1]^d$.

It can be proved that X and $X + h$ are measurable functions of the sequences Z and $Z + c$, respectively. This implies that to prove the equivalence of the laws P^X and P^{X+h} it suffices to show that the laws of the sequences Z and $Z + c$ are equivalent measures on the sequence space $\mathbb{R}^\infty$. Now for a fixed i, the squared Hellinger distance (see (8.2)) between the laws of Z_i and $Z_i + c_i$ equals

$$1 - e^{-\frac{1}{8}c_i^2} \leq \tfrac{1}{8}c_i^2.$$

Since $c \in \ell^2$, Theorem 11.13 yields the equivalence of the laws.

The ratio of the densities of Z_i and $Z_i + c_i$ at the point z_i is given by $\exp((z_i c_i - c_i^2/2))$. Therefore, the Radon-Nikodym derivative of the law of $Z + c$ relative to the law of Z at the point $Z = (Z_1, Z_2, \ldots)$ is given by

$$\prod_{i=1}^\infty e^{c_i Z_i - \frac{1}{2}c_i^2} = e^{\sum c_i Z_i - \frac{1}{2}\sum c_i^2}.$$

This completes the proof, since $Uh = \sum c_i Z_i$ and $\sum c_i^2 = \|h\|_{\mathbb{H}}^2$. □

The converse of the Cameron-Martin theorem can be proved as well: if $h \notin \mathbb{H}$, then the laws P^X and P^{X+h} are singular.

11.3 Support and concentration

Recall that the *support* of the centered, continuous Gaussian process X on $[0,1]^d$ is defined to be the smallest closed subset $F \subset C[0,1]^d$ such that $\Pr(X \in F) = 1$ (Definition 2.1). By Exercise 2.2 such a set exists. The following theorem asserts that the support is determined by the RKHS $\mathbb{H}$ of the process.

Theorem 11.4 (Kallianpur) *The support of the process X is the closure of its RKHS $\mathbb{H}$ in $C[0,1]^d$.*

Proof Observe that by definition, $f \in C[0,1]^d$ is in the support F of X if and only if for all $\epsilon > 0$, it holds that $\Pr(\|X - f\|_\infty < \epsilon) > 0$.

Let Y be an independent copy of X and $\epsilon > 0$. Since $C[0,1]^d$ is countable union of balls of radius ϵ, there exists an $f \in C[0,1]^d$ such that $\Pr(\|X - f\|_\infty < \epsilon) > 0$. By the triangle inequality and independence,

$$\Pr(\|X - Y\|_\infty < 2\epsilon) \geq \Pr(\|X - f\|_\infty < \epsilon) \Pr(\|Y - f\|_\infty < \epsilon) > 0.$$

But the process $(X - Y)/\sqrt{2}$ has the same law as X (check!), hence the preceding shows that for all $\epsilon > 0$, $\Pr(\|X\|_\infty < \epsilon) > 0$. By the Cameron-Martin theorem it follows that every element of $\mathbb{H}$ belongs to the support. Since the support is closed by definition, the closure of $\mathbb{H}$ is included in the support as well.

To show that $\mathbb{H}$ is dense in the support F, let b^* be a bounded linear functional on $C[0,1]^d$ that vanishes on $\mathbb{H}$. By Hahn-Banach (see Theorem 11.12) we are done once we have shown that b^* vanishes on F. The random variable b^*X belongs to the first chaos $\mathscr{L}$ and hence $h(t) = \mathrm{E}X_t b^*X$ belongs to $\mathbb{H}$. By the fact that b^* vanishes on $\mathbb{H}$ and linearity we have $0 = b^*h = b^*(\mathrm{E}X.b^*X) = \mathrm{E}(b^*X)^2$, hence $b^*X = 0$ on an event Ω' with probability 1. Now take $f \in F$. Then for every $n \in \mathbb{N}$ we have

$$\Pr(\Omega' \cap \{\omega \colon \|X(\omega) - f\|_\infty < \epsilon\}) > 0,$$

hence f is the uniform limit of sample paths of $X(\omega_n)$ of X for $\omega_n \in \Omega'$. By definition of Ω', we have $b^*X(\omega_n) = 0$ for every n. By continuity of b^*, we conclude that $b^*f = 0$ as well. □

The theorem implies that for $\mathbb{H}_1$ the unit ball of the RKHS, $\mathbb{B}_1$ the unit ball of $C[0,1]^d$ and $\epsilon > 0$, it holds that

$$\Pr(X \in M\mathbb{H}_1 + \epsilon\mathbb{B}_1) \to 1$$

as $M \to \infty$. The following theorem refines this considerably by quantifying the amount mass that the law of X places on neighborhoods of RKHS balls.

We denote the standard normal distribution function by Φ, that is, $\Phi(x) = \Pr(Z \leq x)$, for Z a standard normal random variable.

Theorem 11.5 (Borell-Sudakov) *For all $\epsilon > 0$ and $M \geq 0$,*

$$\Pr(X \in M\mathbb{H}_1 + \epsilon\mathbb{B}_1) \geq \Phi(\Phi^{-1}(\Pr(\|X\|_\infty < \epsilon)) + M).$$

For $x \geq 0$, we have the inequality $1 - \Phi(x) \leq \exp(-x^2/2)$. Hence the Borell-Sudakov theorem implies in particular that for fixed $\epsilon > 0$ and large enough $M > 0$,

$$\Pr(X \notin M\mathbb{H}_1 + \epsilon\mathbb{B}_1) \lesssim e^{-\frac{1}{2}M^2}.$$

Hence if M is only moderately large, the bulk of the mass of the law of X lies in the set $M\mathbb{H}_1 + \epsilon\mathbb{B}_1$. This should be viewed as the infinite-dimensional analogue of the fact that for a d-dimensional $N_d(0, \Sigma)$-distribution, the bulk of the mass lies in the ellipsoid $\{x \in \mathbb{R}^d \colon x^T\Sigma^{-1}x \leq M\}$ for M large enough.

11.4 Small ball probabilities

We saw in the preceding section that for a centered, continuous Gaussian process X and $\epsilon > 0$, the *small ball probability* $\Pr(\|X\|_\infty < \epsilon)$ is strictly positive. The fact that the process X is centered suggests that for a fixed $\epsilon > 0$, the probability $\Pr(\|X - f\|_\infty < \epsilon)$ of a ball centered at some function f is maximal for $f = 0$. It can be proved that this is indeed the case (this is the content of a result called Anderson's lemma). The following results describe how the probability decreases when the ball is centered at a non-zero function.

Lemma 11.6 *For $h \in \mathbb{H}$ and $\epsilon > 0$,*

$$\Pr(\|X - h\|_\infty < \epsilon) \geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \Pr(\|X\|_\infty < \epsilon).$$

Proof Since X and $-X$ have the same distribution, $\Pr(\|X + h\| < \epsilon) = \Pr(\|X - h\| < \epsilon)$. By the Cameron-Martin formula,

$$\Pr(\|X + h\| < \epsilon) = \mathrm{E} e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1_{\|X\| < \epsilon}.$$

This is true with $-h$ in the place of h as well. Combining these two facts we get

$$\begin{aligned} \Pr(\|X - h\| < \epsilon) &= \frac{1}{2}\mathrm{E} e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1_{\|X\| < \epsilon} + \frac{1}{2}\mathrm{E} e^{U(-h) - \frac{1}{2}\|-h\|_{\mathbb{H}}^2} 1_{\|X\| < \epsilon} \\ &= e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \mathrm{E} \cosh(Uh) 1_{\|X\| < \epsilon}, \end{aligned}$$

where $\cosh(x) = (\exp(x) + \exp(-x))/2$. The proof is finished by noting that $\cosh(x) \geq 1$ for all x. □

Any $f \in C[0,1]^d$ in the support of X can be uniformly approximated by an element of the RKHS (Theorem 11.4). By the triangle inequality, the preceding lemma implies that for every such f and $\epsilon > 0$,

$$-\log \Pr(\|X - f\|_\infty < 2\epsilon) \leq \frac{1}{2} \inf_{h \in \mathbb{H}: \|f - h\|_\infty \leq \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pr(\|X\|_\infty < \epsilon)$$

The quantity on the right plays an important role later on and therefore gets a special name.

Definition 11.7 For $f \in C[0,1]^d$ we define the *concentration function* ϕ_f by

$$\phi_f(\epsilon) = \frac{1}{2} \inf_{h \in \mathbb{H}: \|f - h\|_\infty \leq \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pr(\|X\|_\infty < \epsilon) \tag{11.3}$$

In this notation the inequality in the lemma above states that for f in the support of X and $\epsilon > 0$, $-\log \Pr(\|X - f\|_\infty < 2\epsilon) \leq \phi_f(\epsilon)$. The following theorem asserts that we have a reversed inequality as well.

Theorem 11.8 *For $f \in C[0,1]^d$ in the support of X and $\epsilon > 0$,*

$$\phi_f(2\epsilon) \leq -\log \Pr(\|X - f\|_\infty < 2\epsilon) \leq \phi_f(\epsilon).$$

Sketch of proof The proof of the second inequality was given above.

For the first inequality, it can first be observed using convexity considerations that the map $h \mapsto \|h\|_{\mathbb{H}}^2$ attains a minimum on the set $\{h \in \mathbb{H}: \|f - h\|_\infty \leq \epsilon\}$ at some point h_ϵ, say. By the Cameron-Martin formula,

$$\begin{aligned}\Pr(\|X - f\|_\infty < \epsilon) &= \Pr(\|(X - h_\epsilon) - (f - h_\epsilon)\|_\infty < \epsilon) \\ &= \mathrm{E} e^{-Uh_\epsilon - \frac{1}{2}\|h_\epsilon^2\|_{\mathbb{H}}} 1_{\|X-(f-h_\epsilon)\|_\infty < \epsilon}.\end{aligned}$$

Using series expansions for instance, it can be proved that $Uh_\epsilon \geq 0$ on the event $\{\|X - (f - h_\epsilon)\|_\infty < \epsilon\}$. Since X is centered, it then follows that

$$\Pr(\|X - f\|_\infty < \epsilon) \leq e^{-\frac{1}{2}\|h_\epsilon^2\|_{\mathbb{H}}} \Pr(\|X - (f - h_\epsilon)\|_\infty < \epsilon) \leq e^{-\frac{1}{2}\|h_\epsilon^2\|_{\mathbb{H}}} \Pr(\|X\|_\infty < \epsilon).$$

This is exactly the inequality that needs to be derived. □

The asymptotic behavior for $\epsilon \to 0$ of the centered small ball probability $\Pr(\|X\|_\infty < \epsilon)$ on the logarithmic scale turns out to be closely related to the "size" of the RKHS unit ball $\mathbb{H}_1$. More precisely, it is determined by the asymptotic behavior for $\epsilon \to 0$ of the metric entropy $\log N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty)$. (Note that by Theorem 11.2 these entropy numbers are finite.)

To explain the connection between these two quantities, suppose we have N elements $h_1, \ldots, h_N$ of $\mathbb{H}_1$ that are 2ϵ-separated in $C[0,1]^d$. Then the balls of radius ϵ around the h_j are disjoint and hence, by Lemma 11.6,

$$\begin{aligned}\sqrt{e} &\geq \sqrt{e} \sum_{j=1}^N \Pr(\|X - h_j\|_\infty < \epsilon) \\ &\geq \sqrt{e} \sum_{j=1}^N e^{-\frac{1}{2}\|h_j\|_{\mathbb{H}}^2} \Pr(\|X\|_\infty < \epsilon) \\ &\geq N \Pr(\|X\|_\infty < \epsilon).\end{aligned}$$

Since the maximal number of ϵ-separated points in $\mathbb{H}_1$ is bounded from below by the ϵ-covering number of $\mathbb{H}_1$ (check!), we see that large metric entropy of $\mathbb{H}_1$ implies a small probability $\mathbb{P}(\|X\|_\infty < \epsilon)$ and vice versa.

More careful analysis allows to establish a much more precise relation between the asymptotic behavior of the small ball probability and that of the metric entropy of the RKHS unit ball.

Theorem 11.9 *We have the following equivalences of asymptotic behaviors for $\epsilon \to 0$:*

$$\log N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty) \asymp \epsilon^{-\frac{2\alpha}{2+\alpha}} \iff -\log \Pr(\|X\|_\infty < \epsilon) \asymp \epsilon^{-\alpha} \tag{11.4}$$

$$\log N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty) \asymp \log^\gamma \frac{1}{\epsilon} \iff -\log \Pr(\|X\|_\infty < \epsilon) \asymp \log^\gamma \frac{1}{\epsilon}. \tag{11.5}$$

11.5 Examples

11.5.1 Brownian motion

Let $W = (W_t: t \in [0,1])$ be a Brownian motion. The first chaos of the process is given by the collection of Wiener integrals $\mathscr{L} = \{\int_0^1 f(u)\, dW_u: f \in L^2[0,1]\}$ (see Exercise 11.1).

For $f \in L^2[0,1]$ and $t \geq 0$, the isometry (10.1) implies that

$$\mathrm{E} W_t \int_0^1 f(u)\, dW_u = \int_0^t f(u)\, du.$$

Hence, the RKHS of W is the space

$$\mathbb{H} = \{t \mapsto \int_0^t f(u)\, du \colon f \in L^2[0,1]\}$$

of absolutely continuous functions starting in 0, with square integrable derivatives. This is the so-called *Cameron-Martin space*. Using (10.1) again we see that the RKHS inner product is given by

$$\left\langle t \mapsto \int_0^t f(u)\, du, t \mapsto \int_0^t g(u)\, du \right\rangle_{\mathbb{H}} = \langle f, g \rangle_2 .$$

Using this explicit description of the RKHS it can be seen that the support of the Brownian motion is the space $C_0[0,1]$ of all continuous functions f on $[0,1]$ satisfying $f(0) = 0$ (Exercise 11.4).

Note that the RKHS consists of functions with “regularity” 1, i.e. the functions are smoother than the sample paths of the process W. The unit ball $\mathbb{H}_1$ of the RKHS is a Sobolev-type ball of regularity 1 of which it is known that the metric entropy is of the order ϵ^{-1} (compare with Lemma 8.9). According to the relation (11.4) this implies that for $\epsilon \to 0$ the small ball probabilities of the Brownian motion behave like

$$-\log \Pr(\|W\|_\infty < \epsilon) \asymp \epsilon^{-2}.$$

In fact, this probability estimate can also be proved directly, for instance using known facts about the distribution of first time that $|W|$ hits the level ϵ. Then using (11.4) in the other direction we then find the entropy estimate $\log N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty) \asymp \epsilon^{-1}$.

11.5.2 Stationary processes

The RKHS of a stationary process can be described in spectral terms. Let $X = (X_t \colon t \in [0,1]^d)$ be a centered, continuous, stationary Gaussian process with spectral measure μ. As before, we define $e_t \colon \mathbb{R}^d \to \mathbb{C}$ by $e_t(\lambda) = \exp(i\, \langle \lambda, t \rangle)$.

Lemma 11.10 *The RKHS of X is the set of (real parts of) the functions (from $[0,1]^d$ to $\mathbb{C}$)*

$$t \mapsto \int e^{i\langle \lambda, t \rangle} \psi(\lambda)\, \mu(d\lambda),$$

where ψ runs through the complex Hilbert space $L^2(\mu)$. The RKHS-norm of the displayed function equals the norm in $L^2(\mu)$ of the projection of ψ on the closed linear span of the set of functions $\{e_t \colon t \in [0,1]^d\}$ (or, equivalently, the infimum of $\|\psi\|_{L^2(\mu)}$ over all functions ψ giving the same function in the preceding display).

Proof By the spectral isometry (10.3), the first chaos $\mathscr{L}$ of X is isometric to the space of functions $\mathscr{L}' \subset L^2(\mu)$ defined as the closure in $L^2(\mu)$ of the linear span of the functions $\{e_t \colon t \in [0,1]^d\}$. Since every element of $\mathbb{H}$ is of the form $t \mapsto \mathrm{E} X_t L$ for $L \in \mathscr{L}$, using the spectral isometry again shows that every element of $\mathbb{H}$ is of the form $t \mapsto \langle e_t, \psi \rangle_{L^2(\mu)}$, for $\psi \in \mathscr{L}'$, and the RKHS-norm of such a function is given by the $L^2(\mu)$-norm of ψ.

Now let $P\colon L^2(\mu) \to L^2(\mu)$ be the orthogonal projection onto the closed subspace $\mathscr{L}'$. Then for every $\psi \in L^2(\mu)$, $\langle e_t, \psi \rangle_{L^2(\mu)} = \langle Pe_t, \psi \rangle_{L^2(\mu)} = \langle e_t, P\psi \rangle_{L^2(\mu)}$. Hence $t \mapsto \langle e_t, \psi \rangle_{L^2(\mu)}$ belongs to $\mathbb{H}$ and its RKHS norm is given by the $L^2(\mu)$-norm of the projection $P\psi$. □

11.6 COMPLEMENTS

A collection of functions $\mathscr{F} \subset C[0,1]^d$ is called *uniformly bounded* if $\sup_{f \in \mathscr{F}} \|f\|_\infty < \infty$. It is called *uniformly equicontinuous* is for all $\epsilon > 0$, there exists a $\delta > 0$ such that $\|s - t\| < \delta$ implies that $|f(s) - f(t)| \leq \epsilon$ for all $s, t \in [0,1]^d$ and $f \in \mathscr{F}$.

Theorem 11.11 (Arzelà-Ascoli) *The set $\mathscr{F} \subset C[0,1]^d$ is precompact if and only if it is uniformly bounded and uniformly equicontinuous.*

The following is a version of (or a consequence of) the Hahn-Banach theorem.

Theorem 11.12 *(Hahn-Banach) Let W be a linear subspace of a normed linear space V. If every bounded linear functional on V that vanishes on W, vanishes on the whole space V, then W is dense in V.*

In the following theorem, h denotes the Hellinger distance between densities (see (8.2)).

Theorem 11.13 (Kakutani) *Let $X = (X_1, X_2, \ldots)$ and $Y = (Y_1, Y_2, \ldots)$ be two sequences of independent random variables. Assume X_i has a positive density f_i with respect to a dominating measure μ, and Y_i has a positive density g_i with respect to μ. Then the laws of the sequences X and Y are equivalent probability measures on $\mathbb{R}^\infty$ if and only if*

$$\sum_{i=1}^\infty h^2(f_i, g_i) < \infty.$$

If the laws are not equivalent, they are mutually singular.

Exercises

11.1 Prove that the first chaos of the Brownian motion $W = (W_t\colon t \in [0,1])$ can be identified with the collection of Wiener integrals $\{\int_0^1 f(u)\, dW_u\colon f \in L^2[0,1]\}$.

11.2 Prove that a Gaussian process with continuous sample paths is mean-square continuous.

11.3 Prove that the RKHS is a separable Hilbert space.

11.4 Determine the support of the Brownian motion indexed by $[0,1]$.

11.5 Determine the RKHS of integrated Brownian motion.

12

Contraction rates for Gaussian process priors

12.1 A general theorem for Gaussian processes

Let $W = (W_t : t \in [0,1]^d)$ be a centered, continuous Gaussian process with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$. The process W can be used to define prior distributions on functional parameters in various types of nonparametric statistical models. In a regression context for instance the law of the process W itself can serve as a prior on an regression function, cf. Section 10.4. In other settings it is sensible to first transform the process. If a prior on a density is required for instance, the law of the transformed Gaussian process

$$t \mapsto \frac{e^{W_t}}{\int_{[0,1]^d} e^{W_s}\,ds} \tag{12.1}$$

may be used. Such priors are sometimes called *logistic Gaussian priors*.

In the density estimation setting, we have a general rate of contraction theorem for posterior distribution, Theorem 9.3. (Such results exists for other settings as well, including regression.) The following general theorem will allow us to derive contraction rates for priors based on Gaussian processes. If will turn out that the rate ϵ_n is determined by the concentration function of the process W, as defined in Definition 11.7.

Theorem 12.1 *Let w_0 be contained in the support of W. For any sequence ϵ_n of positive numbers satisfying*

$$\phi_{w_0}(\epsilon_n) \le n\epsilon_n^2 \tag{12.2}$$

and any constant $C > 1$ such that $e^{-Cn\epsilon_n^2} < 1/2$, there exists a sequence of Borel measurable sets $B_n \subset C[0,1]^d$ such that

$$\log N(3\epsilon_n, B_n, \|\cdot\|_\infty) \le 6Cn\epsilon_n^2, \tag{12.3}$$

$$\Pr(W \notin B_n) \le e^{-Cn\epsilon_n^2}, \tag{12.4}$$

$$\Pr\big(\|W - w_0\|_\infty < 2\epsilon_n\big) \ge e^{-n\epsilon_n^2}. \tag{12.5}$$

Proof Inequality (12.5) is an immediate consequence of (12.2) and Theorem 11.8. We need to prove existence of the sets B_n such that the first and third inequalities hold.

For $\mathbb{B}_1$ and $\mathbb{H}_1$ the unit balls in the space $C[0,1]^d$ and the RKHS $\mathbb{H}$, respectively, and $M_n \to \infty$ a sequence of constants to be determined later, set

$$B_n = \epsilon_n\mathbb{B}_1 + M_n\mathbb{H}_1.$$

By the Borell-Sudakov inequality (see Theorem 11.5), it follows that

$$\Pr(W \notin B_n) \le 1 - \Phi(\alpha_n + M_n),$$

for Φ the distribution function of the standard normal distribution and α_n determined by

$$\Phi(\alpha_n) = \Pr(W \in \epsilon_n \mathbb{B}_1) = e^{-\phi_0(\epsilon_n)}.$$

For $C > 1$, set

$$M_n = -2\Phi^{-1}\big(e^{-Cn\epsilon_n^2}\big).$$

Because $\phi_0(\epsilon_n) \le \phi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ by assumption (12.2), and $C > 1$, we have that $\alpha_n \ge -\frac{1}{2}M_n$ so that $\alpha_n + M_n \ge \frac{1}{2}M_n$ and hence

$$\Pr(W \notin B_n) \le 1 - \Phi(\tfrac{1}{2}M_n) = e^{-Cn\epsilon_n^2}.$$

We conclude that (12.4) is satisfied.

If $h_1, \ldots, h_N$ are contained in $M_n\mathbb{H}_1$ and $2\epsilon_n$-separated for the norm $\|\cdot\|_\infty$, then the $\|\cdot\|_\infty$-balls $h_j + \epsilon_n\mathbb{B}_1$ of radius ϵ_n around these points are disjoint and hence

$$\begin{aligned} 1 &\ge \sum_{j=1}^N \Pr(W \in h_j + \epsilon_n\mathbb{B}_1) \\ &\ge \sum_{j=1}^N e^{-\frac{1}{2}\|h_j\|_{\mathbb{H}}^2} \Pr(W \in \epsilon_n\mathbb{B}_1) \\ &\ge N e^{-\frac{1}{2}M_n^2} e^{-\phi_0(\epsilon_n)}, \end{aligned}$$

where the second inequality follows from the Cameron-Martin theorem, Theorem 11.3. If the $2\epsilon_n$-net $h_1, \ldots, h_N$ is maximal in the set $M_n\mathbb{H}_1$, then the balls $h_j + 2\epsilon_n\mathbb{B}_1$ cover $M_n\mathbb{H}_1$. It follows that

$$N\big(2\epsilon_n, M_n\mathbb{H}_1, \|\cdot\|_\infty\big) \le N \le e^{\frac{1}{2}M_n^2} e^{\phi_0(\epsilon_n)}.$$

By its definition any point of the set B_n is within distance ϵ_n of some point of $M_n\mathbb{H}_1$. This implies that

$$\begin{aligned} \log N\big(3\epsilon_n, B_n, \|\cdot\|\big) &\le \log N\big(2\epsilon_n, M_n\mathbb{H}_1, \|\cdot\|\big) \\ &\le \tfrac{1}{2}M_n^2 + \phi_0(\epsilon_n) \\ &\le 5Cn\epsilon_n^2 + \phi_0(\epsilon_n), \end{aligned}$$

by the definition of M_n if $e^{-\frac{1}{2}Cn\epsilon_n^2} < 1$, because $\Phi^{-1}(y) \ge -\sqrt{5/2\log(1/y)}$ and is negative for every $y \in (0, 1/2)$. Since $\phi_0(\epsilon_n) \le \phi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ this completes the proof of the theorem. □

We note that since the concentration function (11.3) is the sum of two functions that are decreasing near 0, condition (12.2) can be replaced by the two separate conditions

$$\tfrac{1}{2}\int_{h\in\mathbb{H}:\|h-w_0\|_\infty \le \tilde\epsilon_n} \|h\|_{\mathbb{H}}^2 \le n\tilde\epsilon_n^2,$$

$$-\log \Pr(\|W\|_\infty < \bar\epsilon_n) \le n\bar\epsilon_n^2.$$

Inequalities (12.3)–(12.5) will then be fulfilled for the maximum $\epsilon_n = \tilde\epsilon_n \vee \bar\epsilon_n$. The second condition only involves the process W, not the function w_0. The first condition measures how well the function w_0 can be approximated by elements of the RKHS of W.

12.2 Density estimation with logistic Gaussian process priors

The inequalities (12.3)–(12.5) in the conclusion of Theorem 12.1 are obviously closely related to the conditions of the general rate of contraction result given by Theorem 9.3. To be able to apply Theorem 12.1 in the context of i.i.d. density estimation with logistic Gaussian process priors we have to relate the various statistical "distances" between densities of the form (12.1) to the uniform distance between sample paths of the process W.

For $w\colon [0,1]^d \to \mathbb{R}$ a continuous function, define the positive density p_w on $[0,1]^d$ by

$$p_w(t) \mapsto \frac{e^{w(t)}}{\int_{[0,1]^d} e^{w(s)}\,ds}, \qquad t \in [0,1]^d. \tag{12.6}$$

Recall the definitions of the distance measures h, K and V in (8.2) and (9.1).

Lemma 12.2 *For any measurable functions $v, w\colon [0,1]^d \to \mathbb{R}$,*

- $h(p_v, p_w) \le \|v-w\|_\infty\, e^{\|v-w\|_\infty/2}$.
- $K(p_v, p_w) \lesssim \|v-w\|_\infty^2\, e^{\|v-w\|_\infty}(1+\|v-w\|_\infty)$.
- $V(p_v, p_w) \lesssim \|v-w\|_\infty^2\, e^{\|v-w\|_\infty}(1+\|v-w\|_\infty)^2$.

Proof The triangle inequality and simple algebra give

$$\begin{aligned} h(p_v,p_w) &= \Big\| \frac{e^{v/2}}{\|e^{v/2}\|_2} - \frac{e^{w/2}}{\|e^{w/2}\|_2} \Big\|_2 = \Big\| \frac{e^{w/2}-e^{v/2}}{\|e^{w/2}\|_2} + e^{v/2}\Big(\frac{1}{\|e^{w/2}\|_2} - \frac{1}{\|e^{v/2}\|_2}\Big)\Big\|_2 \\ &\le 2\frac{\|e^{v/2}-e^{w/2}\|_2}{\|e^{w/2}\|_2}. \end{aligned}$$

Because $|e^{v/2}-e^{w/2}| = e^{w/2}|e^{v/2-w/2}-1| \le e^{w/2}e^{|v-w|/2}|v-w|/2$ for any $v, w \in \mathbb{R}$, the square of the right side is bounded by

$$\frac{\int e^w e^{|v-w|}|v-w|^2\,dt}{\int e^w\,dt} \le e^{\|v-w\|_\infty}\|v-w\|_\infty^2.$$

This proves the first assertion of the lemma.

We derive the other assertions from the first one. Because $w - \|v-w\|_\infty \le v \le w + \|v-w\|_\infty$,

$$\int e^w\,dt\, e^{-\|v-w\|_\infty} \le \int e^v\,dt \le \int e^w\,dt\, e^{\|v-w\|_\infty}.$$

Taking logarithms across we see that

$$-\|v-w\|_\infty \le \log\frac{\int e^v\,dt}{\int e^w\,dt} \le \|v-w\|_\infty.$$

Therefore

$$\Big\|\log \frac{p_v}{p_w}\Big\|_\infty = \Big\|v - w - \log \frac{\int e^v\,dt}{\int e^w\,dt}\Big\|_\infty \le 2\|v-w\|_\infty.$$

The second and third inequalities now follow from the first by Lemma 9.9. □

Now suppose that we observe a sample $X_1, X_2, \ldots, X_n$ from a positive, continuous density p_0 on $[0,1]^d$. The following theorem shows that for the logistic Gaussian process prior defined as the law of the transformed process (12.1), the rate of posterior convergence is determined by the condition (12.2) on the concentration function, with $w_0 = \log p_0$.

Theorem 12.3 *Let W be a centered, continuous Gaussian process on $C[0,1]^d$ on let the prior Π be the law of of p_W. Suppose that $w_0 = \log p_0$ is contained in the support of W. Then the posterior distribution satisfies*

$$\mathrm{E}_0\Pi_n\big(p{:}\, h(p,p_0) > M\epsilon_n \,|\, X_1,\ldots,X_n\big) \to 0$$

for any sufficiently large constant M and ϵ_n given by (12.2).

Proof We choose the set $\mathcal{P}_n$ in Theorem 9.3 equal to the set $\mathcal{P}_n = \{p_w{:}\, w \in B_n\}$ for $B_n \subset C[0,1]^d$ the measurable set as in Theorem 12.1, with C a large constant. In view of the first inequality of Lemma 12.2 for sufficiently large n the $4\epsilon_n$-entropy of $\mathcal{P}_n$ relative to the Hellinger distance is bounded above by the $3\epsilon_n$-entropy of the set B_n relative to the uniform distance, which is bounded by $6Cn\epsilon_n^2$ by Theorem 12.1. The prior probability $\Pi(\mathcal{P}_n^c)$ outside the set $\mathcal{P}_n$ as in (9.4) is bounded by the probability of the event $\{W \notin B_n\}$, which is bounded by $e^{-Cn\epsilon_n^2}$ by Theorem 12.1. Finally, by the second and third inequalities of Lemma 12.2 the prior probability as in (9.3), but with ϵ_n replaced by a multiple of ϵ_n, is bounded above by the probability of the event $\{\|W - w_0\|_\infty < 2\epsilon_n\}$, which is bounded above by $e^{-n\epsilon_n^2}$ by Theorem 12.1. □

12.3 Example: density estimation with Riemann-Liouville priors

12.3.1 Brownian motion

Let $W = (W_t{:}\, t \in [0,1])$ be a Brownian motion with standard normal initial distribution. That is, $W_t = Z + B_t$, where B is a standard Brownian motion and $Z \sim N(0,1)$, independent of B. Then we have

$$-\log \Pr(\|W\|_\infty < 2\epsilon) \le -\log \Pr(|Z| < \epsilon) - \log \Pr(\|B\|_\infty < \epsilon) \lesssim \epsilon^{-2}$$

(see Section 11.5.1). It can be verified that the RKHS of W is given by $\mathbb{H} = \{t \mapsto a + \int_0^t f(s)\,ds{:}\, a \in \mathbb{R}, f \in L^2[0,1]\}$ (see Exercise 12.1). In other words, it is the space of all functions on $[0,1]$ with square-integrable derivatives, without a restriction at 0. Moreover, the RKHS-norm of $h \in \mathbb{H}$ is given by $\|h\|_{\mathbb{H}}^2 = h^2(0) + \|h'\|_2^2$. Note that this implies that the support of W is the whole space $C[0,1]$ (cf. Theorem 11.4).

To bound the infimum term in the concentration function we assume a certain degree of regularity on w_0. Specifically, suppose that $w_0 \in C^\beta[0,1]$ for $\beta \in (0,1]$. Let ϕ be the standard normal probability density and put $\phi_\sigma(x) = \sigma^{-1}\phi(x/\sigma)$ for $\sigma > 0$. We are going to approximate the function w_0 by convolutions of the form $\phi_\sigma * w_0$ which are defined by

$$(\phi_\sigma * w_0)(t) = \int w_0(s)\phi_\sigma(t-s)\,ds,$$

where we extended w_0 to the whole real line without increasing its C^β-norm (this is always possible). Note that since w_0 is β-Hölder by assumption, we have

$$\begin{aligned}|(\phi_\sigma * w_0)(t) - w_0(t)| &= \Big|\int (w_0(t-s) - w(t))\phi_\sigma(s)\,ds\Big|\\ &\le \|w_0\|_\beta \int |s|^\beta \phi_\sigma(s)\,ds\\ &= \|w_0\|_\beta \sigma^\beta \int |s|^\beta \phi(s)\,ds,\end{aligned}$$

hence $\|\phi_\sigma * w_0 - w_0\|_\infty \le C\sigma^\beta$ for some $C > 0$. Since ϕ_σ is smooth the function $\phi_\sigma * w_0$ is differentiable and using the fact that $\int \phi'(t)\,dt = 0$ we see that

$$\begin{aligned}|(\phi_\sigma * w_0)'(t)| &= \Big|\int (w_0(t-s) - w_0(t))\phi'_\sigma(s)\,ds\Big|\\ &\le \|w_0\|_\beta \int |s|^\beta |\phi'_\sigma(s)|\,ds\\ &= \|w_0\|_\beta \sigma^{\beta-1} \int |s|^\beta |\phi'(s)|\,ds.\end{aligned}$$

Since we also have $|(\phi_\sigma * w_0)(0)| \le \|w_0\|_\infty$ it follows that (the restriction to $[0,1]$ of) $\phi_\sigma * w_0$ belongs to the RKHS $\mathbb{H}$, and $\|\phi_\sigma * w_0\|^2_{\mathbb{H}} \le C'\sigma^{2\beta-2}$ for some $C' > 0$. Choosing $\sigma \asymp \epsilon^{1/\beta}$ above we arrive at the bound

$$\inf_{h\in\mathbb{H}:\|h-w_0\|\le\epsilon} \|h\|^2_{\mathbb{H}} \lesssim \epsilon^{\frac{2\beta-2}{\beta}}.$$

Now observe that the bound that we have for the concentration function depends on the regularity β of w_0. If $\beta \le 1/2$, then the bound on the infimum dominates and we have $\phi_{w_0}(\epsilon) \lesssim \epsilon^{\frac{2\beta-2}{\beta}}$. The inequality (12.2) is then fulfilled for $\epsilon_n \asymp n^{-\beta/2}$. If $\beta \ge 1/2$, then the small ball term dominates. Then we have $\phi_{w_0}(\epsilon) \lesssim \epsilon^{-2}$, yielding $\epsilon_n = n^{-1/4}$.

So by Theorem 12.3, using the law of the transformed Brownian motion W as a prior in the density estimation problem leads to posterior that contracts around the true density at the rate $n^{-(1/2\wedge\beta)/2}$, where β is the degree of Hölder regularity of the true log-density. It is well known in statistics that the "best" rate at which a β-regular density can be estimated is $n^{-\beta/(1+2\beta)}$. (Such statements are for instance made precise by so-called *minimax theorems*.) We only achieve this rate with the Brownian motion prior if $\beta = 1/2$. Note that the level $1/2$ is precisely the "degree of regularity" of the Brownian motion sample paths. So we see that in this case we achieve an optimal convergence rate if the regularity of the prior matches the regularity of the unknown density that we are estimating.

12.3.2 General Riemann-Liouville processes

The observations at the end of the preceding subsection suggest that if we want to estimate a function with some arbitrary degree of smoothness β at the optimal rate $n^{-\beta/(1+2\beta)}$, we should perhaps use a prior based on a Gaussian process with the same regularity β.

A candidate is the Riemann-Liouville process R^β considered in Section 10.3.2. Similar to the Brownian motion case considered above we have to slightly modify the process to enlarge its support. The RL process starts at 0, and so do its derivatives, if they exist. This implies that the process can only accurately approximate functions with the same property. To remove this restriction we "release" it at zero by adding an independent polynomial. Specifically, let $\beta > 0$ and let R^β be the RL process with parameter β. Let $\underline{\beta}$ be the largest integer strictly smaller than β and defined the *modified Riemann-Liouville process* W^β by

$$W_t^\beta = \sum_{j=0}^{\underline{\beta}+1} Z_j t^j + R_t^\beta, \qquad t \in [0,1],$$

where the Z_j are independent, standard normal random variable independent of R^β. The following theorem can be proved by analyzing the RKHS of the RL process.

Theorem 12.4 *Let $\beta > 0$. The support of the process W^β is the whole space $C[0,1]$. For $w_0 \in C^\beta[0,1]$, its concentration function ϕ_{w_0} satsfies $\phi_{w_0}(\epsilon) \lesssim \epsilon^{-1/\beta}$ for ϵ small enough.*

The theorem implies that for W^β the modified RL process and $w_0 \in C^\beta[0,1]$, the inequality $\phi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$ is solved for ϵ_n a multiple of $n^{-\beta/(1+2\beta)}$. Hence by Theorem 12.3, using the transformed process W^β as a prior in the density estimation setting yields the optimal rate $n^{-\beta/(1+2\beta)}$ if the true log-density belongs to $C^\beta[0,1]$. So again, we see that we attain an optimal rate of contraction if the regularity of the prior matches the regularity of the truth.

It should be noted that in a realistic situation we typically do not know the regularity of the unknown density that we are trying to estimate. Hence, it is then unclear which prior to use and using one that is too smooth or too rough will lead to sub-optimal rates. This raises the question whether it is possible to construct a prior that achieves the rate $n^{-\beta/(1+2\beta)}$ if the true log-density belongs to $C^\beta[0,1]$, but that does *not* depend on this information about the unknown smoothness. It turns out that this is indeed possible, but the treatment of these so-called *rate-adaptive* procedures is outside the scope of this course.

Exercises

12.1 Determine the RKHS and the RKHS-norm of the Brownian motion with standard normal initial distribution.

13

Efficiency in smooth parametric models

In this lecture we consider estimation in smooth, parametric models and formulate a theorem that characterizes *efficiency*, a notion of optimality in the class of all so-called *regular* estimators. The aim of this lecture is to prepare for the Bernstein-von Mises theorem, which asserts that posterior distributions in smooth parametric models tend to follow efficienct estimators and lead to efficient inferential conclusions.

13.1 Confidence sets and credible sets

Statistical inference can be expressed in many different ways but perhaps the most intuitive and straightforward is the representation in terms of *confidence sets* (in frequentism) or *credible sets* (in Bayesian jargon). Recall that, typically, confidence sets are defined as neighbourhoods of an estimator, of a size or radius that is derived from the quantiles of its *sampling distribution*. The sampling distribution describes the distribution of the estimator relative to the true value of the parameter. That means that sampling distributions provide *coverage probabilities*: for any neighbourhood of the estimator, the sampling distribution tells you the probability that it includes the true parameter. Reversing the reasoning, the analysis often departs from the follwing definition, in which we assume a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for data X.

Definition 13.1 Given a fixed *confidence level* $1 - \alpha$ (for small values of $0 < \alpha < 1$, *e.g.* $\alpha = 0.05$ or 0.01), a *confidence set* $C(X)$ is a model subset, defined in terms of the data X, such that $P_\theta(\theta \in C(X)) \geq 1 - \alpha$, for all $\theta \in \Theta$.

The definition aims for sets that are statistics, (*i.e.* that can be calculated once the data has been realised), with coverage probability greater than or equal to $1-\alpha$. Since there are many sets that satisfy this requirement (note that the entire parameter space always has coverage 1) one strives to find a candidate of maximal informative content, for instance by insisting on minimal Lebesgue measure. An example is a search for intervals of minimal length, in the case where we are considering confidence sets for a one-dimensional parameter based on an estimator that has a unimodal or symmetric distribution around the truth. (More to the point, minimality of the Lebesgue measure of a confidence set is guaranteed if we consider level sets of the Lebesgue density of the sampling distribution.)

Credible sets are motivated conceptually in a very similar way, albeit in Bayesian terms. For the following definition, assume that the model $\mathcal{P}$ (or Θ) is equipped with a prior Π with associated posterior $\Pi(\cdot|X)$.

Definition 13.2 Given a *credible level* $1-\alpha$, a *credible set* $D(X)$ is a model subset defined in terms of the data, that receives posterior probability at or above the credible level: $\Pi(\theta \in D(X)|X) \geq 1-\alpha$.

It is noted here that at the conceptual level, the Bayesian posterior plays a role comparable to that of the frequentist sampling distribution: it is a distribution on the model or on its parameter space, supposedly informative at the inferential level. In the following, we shall not be too strict in Bayesian, subjectivist orthodoxy and interpret the posterior as a frequentist device, in a role very close conceptually to that of the sampling distribution of an estimator used above.

From that perspective, a natural question is: do credible sets and confidence sets have something to do with each other? Since they are conceptually so close, could it be that they are close also mathematically? (At least, when explained as frequentist devices)? In this lecture, we discuss efficiency of (point-)estimation culminating in the so-called convolution theorem, which provides a notion of asymptotic optimality at the inferential level (for the class of all regular estimators). In the next lecture, we switch to the Bayesian leg of the story and show that a correspondence between confidence sets and credible sets does exist asymptotically in smooth, parametric models.

The so-called Bernstein-von Mises theorem (Le Cam (1990)) (see theorem 14.1 below) not only demonstrates asymptotic equivalence of credible sets and confidence sets, it also shows that the relevant sets are optimal in the sense that they are associated with so-called *efficient* estimators. Essential to the development of the optimality theory are two concepts: differentiability of the model and regularity of the estimator. Combined, these two properties lead to a notion of optimality comparable to estimators that achieve optimality within the family of unbiased estimators in the Cramér-Rao sense.

13.2 Optimality in smooth, parametric estimation problems

The concept of efficiency has its origin in Fisher's 1920's claim of asymptotic optimality of the maximum-likelihood estimator in differentiable parametric models (Fisher (1959)). Here, optimality of the ML estimate means that they are consistent, achieve optimal $n^{-1/2}$ rate of convergence and possessed a asymptotic sampling distribution of minimal variance. In 1930's and –40's, Fisher's ideas on optimality in differentiable models were sharpened and elaborated upon (see, *e.g.* Cramér (1946)). To illustrate, consider the following classical result from M-estimation (which can be found as theorem 5.23 in in van der Vaart (1998)).

Theorem 13.3 *Let Θ be open in $\mathbb{R}^k$ and assume that $\mathscr{P}$ is characterized by densities $p_\theta\colon \mathscr{X} \to \mathbb{R}$ such that $\theta \mapsto \log p_\theta(x)$ is differentiable at θ_0 for all $x \in \mathscr{X}$, with derivative $\dot{\ell}_\theta(x)$. Assume that there exists a function $\dot{\ell}\colon \mathscr{X} \to \mathbb{R}$ such that $P_0\dot{\ell}^2 < \infty$ and*

$$\left|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\right| \leq \dot{\ell}(x)\, \|\theta_1 - \theta_2\|,$$

for all θ_1, θ_2 in an open neighbourhood of θ_0. Furthermore, assume that $\theta \mapsto P_\theta \log p_\theta$ has a second-order Taylor expansion around θ_0 of the form,

$$P_{\theta_0} \log p_\theta = P_{\theta_0} \log p_{\theta_0} + \tfrac{1}{2}(\theta-\theta_0)^T I_{\theta_0}(\theta-\theta_0) + o(\|\theta-\theta_0\|^2),$$

with non-singular I_{θ_0}. If $(\hat{\theta}_n)$ is a sequence satisfying,

$$\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_{\theta \in \Theta} \mathbb{P}_n \log p_{\hat{\theta}_n} - o_{P_{\theta_0}}(n^{-1}),$$

such that $\hat{\theta}_n \xrightarrow{\theta_0} \theta_0$, then the estimator sequence is asymptotically linear,

$$n^{1/2}(\hat{\theta}_n - \theta_0) = n^{-1/2} \sum_{i=1}^{n} I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1)$$

In particular, $n^{1/2}(\hat{\theta}_n - \theta_0) \overset{\theta_0}{\rightsquigarrow} N(0, I_{\theta_0}^{-1})$.

The last assertion of theorem 13.3 says that the (near-)maximum-likelihood estimators $(\hat{\theta}_n)$ are asymptotically consistent, converge at rate $n^{-1/2}$ and have the inverse Fisher information $I_{\theta_0}^{-1}$ as the covariance matrix for their (normal) limit distribution. At this stage of the discussion, we do not have an argument to show that this asymptotic behaviour is in any sense optimal. Nevertheless, let us take the opportunity to illustrate briefly how asymptotic behaviour translates into inference on θ by considering associated asymptotic confidence sets.

An *asymptotic confidence set* is an approximate confidence set that is derived not from exact sampling distributions, but from approximations implied by limit distributions, *e.g.* from $n^{1/2}(\hat{\theta}_n - \theta_0) \overset{\theta_0}{\rightsquigarrow} N(0, I_{\theta_0}^{-1})$ in the above example. To demonstrate, first suppose that the model is one-dimensional and satisfies the conditions of theorem 13.3. Denoting quantiles of the standard normal distribution by ξ_α, we see from the last assertion of the theorem that:

$$P_{\theta_0}^n\Big(-\xi_\alpha I_{\theta_0}^{1/2} < n^{1/2}(\hat{\theta}_n - \theta_0) \leq \xi_\alpha I_{\theta_0}^{1/2}\Big) \to 1 - 2\alpha,$$

If the Fisher information were known, this would give rise immediately to a confidence interval: the above display implies that,

$$\Big[\,\hat{\theta}_n - n^{-1/2}\,\xi_\alpha\, I_{\theta_0}^{1/2}\,,\, \hat{\theta}_n + n^{-1/2}\,\xi_\alpha\, I_{\theta_0}^{1/2}\,\Big]$$

has coverage probability $1 - 2\alpha$. Since the Fisher information is not known exactly, we substitute an estimator for it, for example the sample variance S_n^2, to arrive at a *studentized* version of the above, which has the same asymptotic coverage and can therefore be used as an asymptotic confidence interval. But we could also have chosen to "plug in" the estimator $\hat{\theta}_n$ for θ_0 in the expression for the Fisher information to arrive at an estimate $I_{\hat{\theta}_n}$. To generalize to higher-dimensional $\Theta \subset \mathbb{R}^k$, recall that if Z has a k-dimensional multivariate normal distribution $N_k(0, \Sigma)$, then $Z^T \Sigma^{-1} Z$ possess a χ^2-distribution with k degrees of freedom. Denoting quantiles of the χ^2-distribution with k degrees of freedom by $\chi^2_{k,\alpha}$, we find that ellipsoids of the form,

$$C_\alpha(X_1, \ldots, X_n) = \big\{\theta \in \Theta \colon n(\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta - \hat{\theta}_n) \leq \chi^2_{k,\alpha}\big\}, \qquad (13.1)$$

have coverage probabilities converging to $1 - \alpha$ and are therefore asymptotic confidence sets.

13.3 Regularity and efficiency

Theorem 13.3 requires a rather large number of smoothness properties of the model: log-densities are required to be differentiable and Lipschitz and the Kullback-Leibler divergence must display a second-order expansion with non-singular second derivative matrix. These sufficient conditions are there to guarantee that the ML estimator displays a property known as *regularity* and the conditions listed are usually referred to as "regularity condtions". The prominence of regularity in the context of optimality questions was not fully appreciated until Hodges discoved an estimator that displayed a property now known as *superefficiency*.

Example 13.4 (Hodges (1951))
Suppose that we estimate a parameter $\theta \in \Theta = \mathbb{R}$ with an estimator sequence (T_n), satisfying limiting behaviour described by $n^{1/2}(T_n - \theta) \overset{\theta}{\rightsquigarrow} L_\theta$ for some laws L_θ, for all $\theta \in \Theta$. In addition, we define a so-called *shrinkage* estimator S_n, by $S_n = T_n$ as long as $|T_n| \geq n^{-1/4}$ and $S_n = 0$ otherwise. The name *shrinkage estimator* arises because any realization of T_n that is close enough to 0 is shrunk to 0 fully. One shows quite easily that S_n has the same asymptotic behaviour as T_n as long as $\theta \neq 0$, *i.e.* $n^{1/2}(S_n - \theta) \overset{\theta}{\rightsquigarrow} L_\theta$ if $\theta \neq 0$. By contrast, if $\theta = 0$, $\epsilon_n(S_n - 0) \overset{0}{\rightsquigarrow} 0$ for *any* rate sequence ϵ_n. In other words, the asymptotic quality of S_n is as good as that of T_n, or strictly better if $\theta = 0$! *(NB: Superefficiency does come at a price, paid in terms of the behaviour of risk functions in neighbourhoods of the point of shrinkage. Furthermore, superefficiency can be achieved on a subset of Lebesgue measure no greater than zero. There is no such thing as a free lunch.)*

So at certain points in the parameter space, Hodges' shrinkage estimators and other superefficient estimators outperform the MLE and other estimators like it asymptotically, while doing equally well for all other points. Superefficiency indicated that Fisher's 1920's claim was false without further refinement and that a comprehensive understanding of optimality in differentiable estimation problems remained elusive.

To resolve the issue and arrive at a sound theory of asymptotic optimality of estimation in differentiable models, we have to introduce two concepts. The first is a concise notion of smoothness that describes local behaviour of likelihood products directly in terms of score functions. The "local" aspect of the definition stems from the n-dependent re-coordinatization in terms of the *local parameter* $h = n^{1/2}(\theta - \theta_0)$. (In the following we assume that the sample is *i.i.d.*, although usually the definition is extended to more general models for the data.)

Definition 13.5 (Local asymptotic normality, Le Cam (1960))
Let $\Theta \subset \mathbb{R}^k$ be open, parametrizing a model $\mathscr{P} = \{P_\theta \colon \theta \in \Theta\}$ that is dominated by a σ-finite measure with densities p_θ. The model is said to be *locally asymptotically normal (LAN)* at θ_0 if, for any converging sequence $h_n \to h$:

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0 + n^{1/2} h_n}}{p_{\theta_0}}(X_i) = h^T \Gamma_{n,\theta_0} - \tfrac{1}{2}\, h^T I_{\theta_0} h + o_{P_{\theta_0}}(1), \tag{13.2}$$

for random vectors Γ_{n,θ_0} such that $\Gamma_{n,\theta_0} \overset{\theta_0}{\rightsquigarrow} N_k(0, I_{\theta_0})$.

Differentiability of the log-density $\theta \mapsto \log p_\theta(x)$ at θ_0 for every x (with score $\dot{\ell}_\theta(x) =$

$(d/d\theta)\log p_\theta(x))$ implies that the model is LAN at θ_0 with,

$$\Gamma_{n,\theta_0} = n^{-1/2}\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i).$$

But local asymptotic normality can be achieved under weaker conditions; best known is the following property, best described as Hadamard differentiability of square-roots of model densities relative to the $L_2(P_0)$ norm.

Definition 13.6 Let $\Theta \subset \mathbb{R}^k$ be open. A model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ that is dominated by a σ-finite measure μ with densities p_θ is said to be differentiable in quadratic mean (DQM) at $\theta_0 \in \Theta$, if there exists a score function $\dot{\ell}_{\theta_0} \in L_2(P_{\theta_0})$ such that:

$$\int \Bigl(p_{\theta_0+h}^{1/2} - p_{\theta_0}^{1/2} - \tfrac{1}{2} h^T\,\dot{\ell}_{\theta_0}\, p_{\theta_0}^{1/2}\Bigr)^2 \,d\mu = o\bigl(\|h\|^2\bigr),$$

as $h \to 0$.

Theorem 7.2 in van der Vaart (1998) shows that a model that is DQM at θ_0 is LAN at θ_0. However, in many situations, it is quite straightforward to demonstrate the LAN property directly.

The second concept is a property that characterizes the class of estimators over which optimality is achieved, in particular excluding Hodges' shrinkage estimators (and all other examples of superefficiency, as becomes clear below). To prepare the definition heuristically, note that, given Hodges' counterexample, it is not enough to have estimators with pointwise convergence to limit laws; we must restrict the behaviour of estimators over ($n^{-1/2}$-)neighbourhoods rather than allow the type of wild variations that make superefficiency possible.

Definition 13.7 Let $\Theta \subset \mathbb{R}^k$ be open. An estimator sequence (T_n) for the parameter θ is said to be *regular* at θ if, for all $h \in \mathbb{R}^k$,

$$n^{1/2}\Bigl(T_n - \bigl(\theta + n^{-1/2}h\bigr)\Bigr) \rightsquigarrow L_\theta, \;\; \text{(under } P_{\theta+n^{-1/2}h}\text{)},$$

i.e. with a limit law independent of h.

So regularity describes the property that convergence of the estimator to a limit law is insensitive to *perturbation* of the parameter of n-dependent size $n^{-1/2}h$. The two properties covered above come together through the following theorem (see theorems 7.10, 8.3 and 8.4 in van der Vaart (1998)), which formulate the foundation for the convolution theorem that follows.

Theorem 13.8 *Let $\Theta \subset \mathbb{R}^k$ be open; let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be LAN at θ_0 with nonsingular Fisher information I_{θ_0}. Let (T_n) be regular estimators in the "localized models" $\{P_{\theta_0+n^{-1/2}h} : h \in \mathbb{R}^k\}$. Then there exists a (randomized) statistic T in the normal location model $\{N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$ such that $T - h \sim L_{\theta_0}$ for all $h \in \mathbb{R}^k$.*

Theorem 13.8 provides every regular estimator sequence with a limit in the form of a statistic in a very simple model in which the only parameter is the location of a normal distribution: the (weak) limit distribution that describes the local asymptotics of the sequence

(T_n) under $P_{\theta_0+n^{-1/2}h}$ *equals* the distribution of T under h, for all $h \in \mathbb{R}^k$. Moreover, regularity of the sequence (T_n) implies that under $N_k(h, I_{\theta_0}^{-1})$, the distribution of T relative to h is independent of h, an invariance usually known as *equivariance-in-law*. (This result is the culmination of a much broader theory of asymptotic behaviour of estimators in sequences of models. This very general theory goes by the name "limits of experiments", see Le Cam (1972) and van der Vaart "Limits of Statistical Experiments" (unpublished).) The class of equivariant-in-law estimators for location in the model $\{N_k(h, I_{\theta_0}^{-1}) \colon h \in \mathbb{R}^k\}$ is fully known: for any equivariant-in-law estimator T for h, there exists a probability distribution M such that $T \sim N_k(h, I_{\theta_0}^{-1}) \star M$. The most straightforward example is $T = X$, for which $M = \delta_0$. This argument gives rise to the following central result in the theory of efficiency.

Theorem 13.9 *(Convolution theorem (Hájek (1970)))*
Let $\Theta \subset \mathbb{R}^k$ be open and let $\{P_\theta \colon \theta \in \Theta\}$ be LAN at θ_0 with non-singular Fisher information I_{θ_0}. Let (T_n) be a regular estimator sequence with limit distribution L_{θ_0}. Then there exists a probability distribution M_{θ_0} such that,

$$L_{\theta_0} = N_k(0, I_{\theta_0}^{-1}) \star M_{\theta_0},$$

In particular, if L_{θ_0} has a covariance matrix Σ_{θ_0}, then $\Sigma_{\theta_0} \geq I_{\theta_0}^{-1}$.

The occurence of the inverse Fisher information is no coincidence: the estimators T are unbiased so they satisfy the Cramér-Rao bound in the limiting model $\{N_k(h, I_{\theta_0}^{-1}) \colon h \in \mathbb{R}^k\}$. Convolution of $N_k(0, I_{\theta_0}^{-1})$ with any distribution M raises its variance unless M is degenerate: the last assertion of the convolution theorem says that, within the class of regular estimates, asymptotic variance is lower-bounded by the inverse Fisher information. A regular estimator that is optimal in this sense, is called *best-regular*. Anderson's lemma broadens the notion of optimality, in the sense that best-regular estimators outperform other regular estimators with respect to a large family of loss functions.

Definition 13.10 A *loss-function* is any $\ell \colon \mathbb{R}^k \to [0, \infty)$; a *subconvex* loss-function is a loss function such that the level sets $\{x \in \mathbb{R}^k \colon \ell(x) \leq c\}$ are closed, convex and symmetric around the origin.

Examples of subconvex loss-functions are many and include, for example, the common choices $\ell(x) = \|x\|^p$, $p \geq 1$.

Lemma 13.11 *(Anderson's lemma)*
For any $k \geq 1$, any subconvex loss function ℓ"$\mathbb{R}^k \to [0, \infty)$, any probability distribution M on $\mathbb{R}^k$ and any k-dimensional covariance matrix Σ,

$$\int \ell \, dN_k(0, \Sigma) \leq \int \ell \, d(N_k(0, \Sigma) \star M).$$

(A proof of Anderson's lemma can be found, for instance, in Ibragimov and Has'minskii (1981).) Finally, we mention the following equivalence, which characterizes efficiency concisely in terms of a weakly converging sequence.

Lemma 13.12 *In a LAN model, estimators (T_n) for θ are best-regular* if and only if *the (T_n) are asymptotically linear, i.e. for all θ in the model,*

$$n^{1/2}(T_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_\theta^{-1} \dot{\ell}_\theta(X_i) + o_{P_\theta}(1). \tag{13.3}$$

The random sequence of differences on the r.h.s. of (13.3) is denoted by Δ_{n,θ_0} below.

Coming back to theorem 13.3, we see that under stated conditions, a consistent sequence of MLE's $(\hat{\theta}_n)$ is best-regular, finally giving substance to Fisher's 1920's claim. Referring to the discussion on confidence sets with which we opened this lecture, we now know that in a LAN model, confidence sets of the form (13.1) based on best-regular estimators $(\hat{\theta}_n)$ satisfy a similar notion of optimality: according to the the convolution theorem, the asymptotic sampling distributions of best-regular estimator sequences are all the same and sharpest among asymptotic sampling distributions for regular estimators. The question remains if we can somehow identify confidence sets and credible intervals; in the next chapter, that identification is made asymptotically.

Exercises

13.1 Assume that $n^{1/2}(\hat{\theta}_n - \theta_0) \sim N(0, I_{\theta_0}^{-1})$. Show that the ellipsoids (13.1) are of minimal Lebesgue measure among all subsets of coverage $1 - \alpha$.

13.2 Consider Hodges' estimators S_n of example 13.4. Show that, for any rate sequence (ϵ_n), $\epsilon_n \downarrow 0$, $\epsilon_n(S_n - 0) \overset{0}{\rightsquigarrow} 0$.

13.3 Let $\Theta = \mathbb{R}$ and let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be the model of Poisson distributions P_θ with means θ. Show that this model is LAN for all θ.

13.4 Let $\Theta = \mathbb{R}$ and let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be the model of normal distributions $N(\theta, 1)$ of unit variance with means θ. Show that this model is LAN for all θ.

13.5 Let f be a Lebesgue density on $\mathbb{R}$ that is symmetric around the origin. Define the model $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$ by densities $f_{\mu,\sigma}(x) = \sigma^{-1} f((x - \mu)/\sigma)$. Show that the Fisher information matrix is diagonal.

14

Le Cam's Bernstein-von Mises theorem

To address the question of efficiency in smooth parametric models from a Bayesian perspective, we turn to the Bernstein-Von Mises theorem. In the literature many different versions of the theorem exist, varying both in (stringency of) conditions and (strength or) form of the assertion. Following Le Cam and Yang (1990) we state the theorem as follows. (For later reference, define a parametric prior to be *thick* at θ_0, if it has a Lebesgue density that is continuous and strictly positive at θ_0.)

Theorem 14.1 *(Bernstein-Von Mises, parametric)*
Assume that $\Theta \subset \mathbb{R}^k$ is open and that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable and dominated. Suppose $X_1, X_2, \ldots$ forms an i.i.d. sample from P_{θ_0} for some $\theta_0 \in \Theta$. Assume that the model is locally asymptotically normal at θ_0 with non-singular Fisher information I_{θ_0}. Furthermore, suppose that, the prior Π_Θ is thick at θ_0 and that for every $\epsilon > 0$, there exists a test sequence (ϕ_n) such that,

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|>\epsilon} P_\theta^n(1-\phi_n) \to 0.$$

Then the posterior distributions converge in total variation,

$$\sup_B \Big| \Pi\big(\, \theta \in B \mid X_1, \ldots, X_n\big) - N_{\hat\theta_n,(nI_{\theta_0})^{-1}}(B) \Big| \to 0,$$

in P_{θ_0}-probability, where $(\hat\theta_n)$ denotes any best-regular estimator sequence.

For any two probability measures P and Q, the total variation norm of their difference is equal to the L_1-norm of the difference of their densities relative to any σ-finite measure that dominates both:

$$\sup_B \big| \, P - Q \, \big| = \int |p-q| \, d\mu, \tag{14.1}$$

where $P, Q \ll \mu$ with $p = dP/d\mu$ and $q = dQ/d\mu$. So if the Lebesgue measure on Θ dominates the posterior, the Bernstein-Von Mises theorem says that the posterior density converges to a normal density in L_1. In figure 14, this type of convergence is demonstrated with a numerical example. Also displayed in figure 14 are the so-called *MAP estimator* (the localtion of maximal posterior density, a popular point-estimator derived from the posterior) and the ML estimator. It is noted that, here, the MLE is efficient so it forms a possible centring sequence for the limiting sequence of normal distributions in the assertion of the Bernstein-Von Mises theorem. Furthermore, it is noted that the posterior concentrates more and more sharply, reflecting the n^{-1}-proportionality of the variance of its limiting sequence

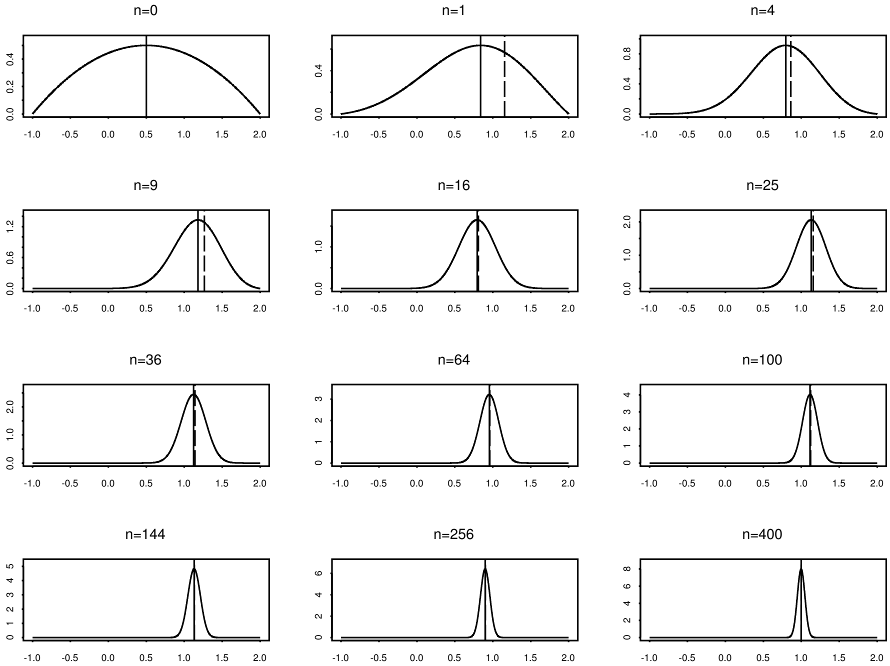

Figure 14.1 Convergence of the posterior density. The samples used for calculation of the posterior distributions consist of n observations; the model consists of all normal distributions with mean between -1 and 2 and variance 1 and has a polynomial prior, shown in the first ($n = 0$) graph. For all sample sizes, the *maximum a posteriori* and maximum likelihood estimators are indicated by a vertical line and a dashed vertical line respectively. (From Kleijn (2003))

of normals. It is perhaps a bit surprising in figure 14 to see limiting normality obtain already at such relatively low values of the sample size n. It cannot be excluded that, in this case, that is a manifestation the normality of the underlying model, but onset of normality of the posterior appears to happen at unexpectedly low values of n also in other smooth, parametric models. It suggests that asymptotic conclusions based on the Bernstein-Von Mises limit accrue validity fairly rapidly, for n in the order of several hundred to several thousand *i.i.d.* replications of the observation.

The uniformity in the assertion of the Bernstein-Von Mises theorem over model subsets B implies that it holds also for model subsets that are random. In particular, given some $0 < \alpha < 1$, it is noted that the smallest sets $C_\alpha(X_1, \ldots, X_n)$ such that,

$$N_{\hat{\theta}_n,(nI_{\theta_0})^{-1}}\bigl(C_\alpha(X_1, \ldots, X_n)\bigr) \geq 1 - \alpha,$$

are ellipsoids of the form (13.1). Since posterior coverage of C_α converges to the *l.h.s.* in the above display, in accordance with the Bernstein-Von Mises limit, we see that the C_α are asymptotic credible sets of posterior coverage $1 - \alpha$. Conversely, any sequence

$(C_n(X_1,\ldots,X_n))$ of (data-dependent) credible sets of coverage $1-\alpha$, is also a sequence of sets that have asymptotic confidence level $1-\alpha$ (where we use that $\hat{\theta}_n$ is best-regular). This gives rise to an identification in smooth, parametric models between inference based on frequentist best-regular point-estimators and inference based on Bayesian posteriors. In a practical sense, it eliminates the need to estimate θ and the Fisher information I_θ at θ to arrive at asymptotic confidence sets, if we have an approximation of the posterior distribution of high enough quality (*e.g.* from MCMC simulation), if we know that the Bernstein-Von Mises theorem holds. In high dimensional parametric models, maximization of the likelihood may be much more costly computationally than generation of a suitable MCMC approximation. As a consequence, the Bernstein-Von Mises theorem has an immediate practical implication of some significance. This point will also hold in semiparametric context, where the comparative advantage is even greater.

Before we continue with the proof of the Bernstein-Von Mises theorem, let is briefly reflect on its conditions: local asymptotic normality and non-singularity of the associated Fisher information are minimal smoothness conditions. They also arise in theorem 13.3 and form the backdrop for any discussion of efficiency. More significant is the required existence of a "consistent" test sequence: what is required is that, asymptotically, we can distinguish P_0 from any complement of a θ-neighbourhood around θ_0 uniformly. One should compare this condition with that of consistency of near-maximizers of the likelihood in theorem 13.3. Apparently, if such a global (rather than $n^{-1/2}$-sized local) consistency guarantee can not be given, likelihood-based methods like ML or Bayesian estimation cannot be trusted to give rise to asymptotic normality (in their respective forms). In the next section, we shall divide the Bernstein-Von Mises theorem in two parts, with a requirement of local $n^{-1/2}$-sized consistency for the posterior in between. In a separate lemma, we show that a score-test can fill in the gap between local and global consistency.

14.1 Proof of the Bernstein-von Mises theorem

We prove the assertion of the Bernstein-Von Mises theorem using a smoothness property that is slightly stronger than necessary, because we shall need a similar formulation in the semiparametric case.

Definition 14.2 We say that a parametric model $\mathscr{P}$ is *stochastically LAN* at θ_0, if the LAN property of definition 13.5 is satisfied for every random sequence (h_n) that is bounded in probability, *i.e.* for all $h_n = O_{P_0}(1)$:

$$\log\prod_{i=1}^{n}\frac{p_{\theta_0+n^{1/2}h_n}}{p_{\theta_0}}(X_i) - h_n^T\Gamma_{n,\theta_0} - \tfrac{1}{2}\,h_n^T I_{\theta_0} h_n = o_{P_{\theta_0}}(1), \tag{14.2}$$

for random vectors Γ_{n,θ_0} such that $\Gamma_{n,\theta_0} \overset{\theta_0}{\rightsquigarrow} N_k(0,I_{\theta_0})$.

Theorem 14.3 *Let the sample $X_1,X_2,\ldots$ be distributed i.i.d.-P_0. Let $\Theta\subset\mathbb{R}^k$ be open, let $\mathscr{P}=\{P_\theta\colon\theta\in\Theta\}$ be stochastically LAN at θ_0 with non-singular Fisher information I_{θ_0} and let the prior Π on Θ be thick. Furthermore, assume that for every sequence of balls $(K_n)\subset\mathbb{R}^d$ with radii $M_n\to\infty$, we have:*

$$\Pi_n\big(\,h\in K_n\;\big|\;X_1,\ldots,X_n\,\big)\xrightarrow{P_0}1. \tag{14.3}$$

Then the sequence of posteriors converges to a sequence of normal distributions in total variation:

$$\sup_B \Bigl| \Pi_n\bigl(h \in B \bigm| X_1, \ldots, X_n \bigr) - N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}(B) \Bigr| \xrightarrow{P_0} 0. \tag{14.4}$$

Proof The proof is split into two parts: in the first part, we prove the assertion conditional on an arbitrary compact set $K \subset \Theta$ and in the second part we use this to prove (14.4). Throughout the proof we denote the posterior for h given $X_1, X_2, \ldots, X_n$ by Π_n and the normal distribution $N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}$ by Φ_n (recall the definition of Δ_{n,θ_0} from lemma 13.12. For $K \subset \mathbb{R}^k$, conditional versions are denoted Π_n^K and Φ_n^K respectively (assuming that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$, of course).

Let $K \subset \Theta$ be a ball centered on 0. For every open neighbourhood $U \subset \Theta$ of θ_0, $\theta_0 + n^{-1/2} K \subset U$ for large enough n. Since θ_0 is an internal point of Θ, we can define, for large enough n, the random functions $f_n \colon K \times K \to \mathbb{R}$ by:

$$f_n(g,h) = \Bigl(1 - \frac{\phi_n(h)}{\phi_n(g)} \frac{s_n(g)}{s_n(h)} \frac{\pi_n(g)}{\pi_n(h)}\Bigr)_+,$$

where $\phi_n \colon K \to \mathbb{R}$ is the Lebesgue density of the (randomly located) distribution $N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}$, $\pi_n \colon K \to \mathbb{R}$ is the Lebesgue density of the prior for the centred and rescaled parameter h and $s_n \colon K \to \mathbb{R}$ equals the likelihood product:

$$s_n(h) = \prod_{i=1}^n \frac{p_{\theta_0 + h/\sqrt{n}}}{p_{\theta_0}}(X_i).$$

Since the model is stochastically LAN by assumption, we have for every random sequence $(h_n) \subset K$:

$$\log s_n(h_n) = h_n \mathbb{G}_n \dot{\ell}_{\theta^*} - \tfrac{1}{2} h_n^T V_{\theta^*} h_n + o_{P_0}(1),$$
$$\log \phi_n(h_n) = -\tfrac{1}{2}(h_n - \Delta_{n,\theta^*})^T V_{\theta^*} (h_n - \Delta_{n,\theta^*}) + \mathit{const}$$

For any two sequences (h_n), $(g_n) \subset K$, $\pi_n(g_n)/\pi_n(h_n) \to 1$ as $n \to \infty$. Combining this with the above display and (15.4), we see that:

$$\begin{aligned}
\log & \frac{\phi_n(h_n)}{\phi_n(g_n)} \frac{s_n(g_n)}{s_n(h_n)} \frac{\pi_n(g_n)}{\pi_n(h_n)} \\
&= -h_n \mathbb{G}_n \dot{\ell}_{\theta^*} + \tfrac{1}{2} h_n^T V_{\theta^*} h_n + g_n \mathbb{G}_n \dot{\ell}_{\theta^*} - \tfrac{1}{2} g_n^T V_{\theta^*} g_n + o_{P_0}(1) \\
&\qquad - \tfrac{1}{2}(h_n - \Delta_{n,\theta^*})^T V_{\theta^*} (h_n - \Delta_{n,\theta^*}) + \tfrac{1}{2}(g_n - \Delta_{n,\theta^*})^T V_{\theta^*} (g_n - \Delta_{n,\theta^*}) \\
&= o_{P_0}(1)
\end{aligned}$$

as $n \to \infty$. Since $x \mapsto (1 - e^x)_+$ is continuous on $\mathbb{R}$, we conclude that for every pair of random sequences $(g_n, h_n) \subset K \times K$:

$$f_n(g_n, h_n) \xrightarrow{P_0} 0, \quad (n \to \infty).$$

For fixed, large enough n, P_0^n-almost-sure continuity of $(g,h) \mapsto \log s_n(g)/s_n(h)$ on $K \times K$ is guaranteed by the stochastic LAN-condition. Each of the locations Δ_{n,θ_0} for Φ_n

is is tight, so $(g,h) \mapsto \phi_n(g)/\phi_n(h)$ is continuous on all of $K \times K$ P_0^n-almost-surely. Continuity (in a neighbourhood of θ_0) and positivity of the prior density guarantee that this holds for $(g,h) \mapsto \pi_n(g)/\pi_n(h)$ as well. We conclude that for large enough n, the random functions f_n are continuous on $K \times K$, P_0^n-almost-surely. Application of lemma 14.5 then leads to the conclusion that,

$$\sup_{g,h\in K} f_n(g,h) \xrightarrow{P_0} 0, \quad (n \to \infty). \tag{14.5}$$

Since K contains a neighbourhood of 0, $\Phi_n(K) > 0)$ is guaranteed. Let Ξ_n denote the event that $\Pi_n(K) > 0$. Let $\eta > 0$ be given and based on that, define the events:

$$\Omega_n = \bigl\{\omega\colon \sup_{g,h\in K} f_n(g,h) \le \eta\bigr\}.$$

Consider the expression (recall that the total-variation norm is bounded by 2):

$$P_0^n \bigl\|\Pi_n^K - \Phi_n^K\bigr\| 1_{\Xi_n} \le P_0^n \bigl\|\Pi_n^K - \Phi_n^K\bigr\| 1_{\Omega_n\cap\Xi_n} + 2P_0^n(\Xi_n - \Omega_n). \tag{14.6}$$

As a result of (14.5) the latter term is $o(1)$ as $n \to \infty$. The remaining term on the *r.h.s.* can be calculated as follows:

$$\begin{aligned}\tfrac12 P_0^n \bigl\|\Pi_n^K - \Phi_n^K\bigr\| 1_{\Omega_n\cap\Xi_n} &= P_0^n \int \Bigl(1 - \frac{d\Phi_n^K}{d\Pi_n^K}\Bigr)_+ d\Pi_n^K \, 1_{\Omega_n\cap\Xi_n}\\ &= P_0^n \int_K \Bigl(1 - \phi_n^K(h) \frac{\int_K s_n(g)\pi_n(g)dg}{s_n(h)\pi_n(h)}\Bigr)_+ d\Pi_n^K(h)\, 1_{\Omega_n\cap\Xi_n}\\ &= P_0^n \int_K \Bigl(1 - \int_K \frac{s_n(g)\pi_n(g)\phi_n^K(h)}{s_n(h)\pi_n(h)\phi_n^K(g)} d\Phi_n^K(g)\Bigr)_+ d\Pi_n^K(h)\, 1_{\Omega_n\cap\Xi_n}.\end{aligned}$$

Note that for all $g,h \in K$ we have $\phi_n^K(h)/\phi_n^K(g) = \phi_n(h)/\phi_n(g)$ since, on K, ϕ_n^K differs from ϕ_n only by a normalisation factor. We use Jensen's inequality (with respect to the Φ_n^K-expectation) for the (convex) function $x \mapsto (1-x)_+$ to derive:

$$\begin{aligned}\tfrac12 P_0^n \bigl\|\Pi_n^K - \Phi_n^K\bigr\| 1_{\Omega_n\cap\Xi_n} &\le P_0^n \int \Bigl(1 - \frac{s_n(g)\pi_n(g)\phi_n(h)}{s_n(h)\pi_n(h)\phi_n(g)}\Bigr)_+ d\Phi_n^K(g)\, d\Pi_n^K(h) 1_{\Omega_n\cap\Xi_n}\\ &\le P_0^n \int \sup_{g,h\in K} f_n(g,h) 1_{\Omega_n\cap\Xi_n} d\Phi_n^K(g)\, d\Pi_n^K(h) \le \eta.\end{aligned}$$

Combination with (14.6) shows that for all compact $K \subset \mathbb{R}^d$ containing a neighbourhood of 0,

$$P_0^n \bigl\|\Pi_n^K - \Phi_n^K\bigr\| 1_{\Xi_n} \to 0.$$

Now let (K_m) be a sequence of balls centred at 0 with radii $M_m \to \infty$. For each $m \ge 1$, the above display holds, so if we choose a sequence of balls (K_n) that traverses the sequence K_m slowly enough, convergence to zero can still be guaranteed. Moreover, the corresponding events $\Xi_n = \{\omega\colon \Pi_n(K_n) > 0\}$ satisfy $P_0^n(\Xi_n) \to 1$ as a result of (14.3). We conclude that there exists a sequence of radii (M_n) such that $M_n \to \infty$ and

$$P_0^n \bigl\|\Pi_n^{K_n} - \Phi_n^{K_n}\bigr\| \to 0, \tag{14.7}$$

(where it is understood that the conditional probabilities on the *l.h.s.* are well-defined on sets of probability growing to one). Combining (14.3) and lemma 14.7, we then use lemma 14.6 to conclude that:

$$P_0^n \big\| \Pi_n - \Phi_n \big\| \to 0,$$

which implies (14.4). □

Aside from a slightly stronger smoothness property in the form of the stochastic LAN condition, theorem 14.3 appears to require more than theorem 14.1, in the sense that it requires posterior consistency at rate $n^{-1/2}$ rather than the (fixed) tests for consistency. The following lemma shows that, assuming smoothness, the latter condition is enough to satisfy the former. Its proof is based on the construction of a score test that fills in the "gap" left between the fixed-alternative tests and the growing alternative $\|\theta - \theta_0\| \geq n^{-1/2} M_n$. However, the proof is long and detailed and it does not have a semiparametric analog. For that reason the proof is given only in the form of a reference.

Lemma 14.4 *Assume that $\Theta \subset \mathbb{R}^k$ is open and that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable and dominated. Assume that the model is locally asymptotically normal at θ_0 with non-singular Fisher information I_{θ_0} and that the prior is thick at θ_0. Furthermore, suppose that there exists a test sequence (ϕ_n) such that,*

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\|\theta - \theta_0\| > \epsilon} P_\theta^n (1 - \phi_n) \to 0.$$

Then the posterior converges at rate $n^{-1/2}$, i.e. for every sequence $M_n \to \infty$,

$$\Pi\big(\|\theta - \theta_0\| \geq n^{-1/2} M_n \;\big|\; X_1, \ldots, X_n \big) \xrightarrow{P_0} 0.$$

Proof This lemma is a well-specified version of theorem 2.2 in Kleijn (2003), which incorporates theorem 2.3 therein, also found as lemma 10.3 in van der Vaart (1998). □

14.2 Three subsidiary lemmas

The proof of theorem 14.3 also makes use of the following three lemmas which are of a more general character then lemma 14.4.

Lemma 14.5 *Let (f_n) be a sequence of random functions $K \to \mathbb{R}$, where K is compact. Assume that for large enough $n \geq 1$, f_n is continuous P_0^n-almost-surely. Then the following are equivalent:*

(i) Uniform convergence in probability:

$$\sup_{h \in K} \big| f_n(h) \big| \xrightarrow{P_0} 0, \quad (n \to \infty),$$

(ii) For any random sequence $(h_n) \subset K$:

$$f_n(h_n) \xrightarrow{P_0} 0, \quad (n \to \infty),$$

Proof *((ii)⇒(i), by contradiction.)* Assume that there exist $\delta, \epsilon > 0$ such that:

$$\limsup_{n\to\infty} P_0\big(\sup_{h\in K} \big|f_n(h)\big| > \delta \big) = \epsilon.$$

Since the functions f_n are continuous P_0-almost-surely, there exists (with P_0-probability one) a sequence $(\tilde{h}_n)$ such that for every $n \geq 1$, $\tilde{h}_n \in K$ and

$$\big|f_n(\tilde{h}_n)\big| = \sup_{h\in K}\big|f_n(h)\big|.$$

Consequently, for this particular random sequence in K, we have:

$$\limsup_{n\to\infty} P_0\Big(\big|f_n(\tilde{h}_n)\big| > \delta\Big) = \epsilon > 0.$$

which contradicts (ii). *((i)⇒(ii).)* Given a random sequence $(h_n) \subset K$, and for every $\delta > 0$,

$$P_0\Big(\sup_{h\in K}\big|f_n(h)\big| > \delta\Big) \geq P_0\Big(\big|f_n(h_n)\big| > \delta\Big).$$

Given (i), the *l.h.s.* converges to zero and hence so does the *r.h.s.*. □

The next lemma shows that given two sequences of probability measures, a sequence of balls that grows fast enough can be used conditionally to calculate the difference in total-variational distance, even when the sequences consist of random measures.

Lemma 14.6 *Let* (Π_n) *and* (Φ_n) *be two sequences of random probability measures on* $\mathbb{R}^k$. *Let* (K_n) *be a sequence of subsets of* $\mathbb{R}^k$ *such that*

$$\Pi_n(\mathbb{R}^k - K_n) \xrightarrow{P_0} 0, \quad \Phi_n(\mathbb{R}^k - K_n) \xrightarrow{P_0} 0. \tag{14.8}$$

Then

$$\big\|\Pi_n - \Phi_n\big\| - \big\|\Pi_n^{K_n} - \Phi_n^{K_n}\big\| \xrightarrow{P_0} 0. \tag{14.9}$$

Proof Let K, a measurable subset of $\mathbb{R}^k$ and $n \geq 1$ be given and assume that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$. Then for any measurable $B \subset \mathbb{R}^k$ we have:

$$\begin{aligned}\big|\Pi_n(B) - \Pi_n^K(B)\big| &= \Big|\Pi_n(B) - \frac{\Pi_n(B\cap K)}{\Pi_n(K)}\Big| \\ &= \big|\Pi_n\big(B\cap(\mathbb{R}^k - K)\big) + \big(1 - \Pi_n(K)^{-1}\big)\Pi_n(B\cap K)\big| \\ &\leq \Pi_n\big(B\cap(\mathbb{R}^k - K)\big) + \Pi_n(\mathbb{R}^k - K)\Pi_n^K(B) \leq 2\Pi_n(\mathbb{R}^k - K).\end{aligned}$$

and hence also:

$$\Big|\big(\Pi_n(B) - \Pi_n^K(B)\big) - \big(\Phi_n(B) - \Phi_n^K(B)\big)\Big| \leq 2\big(\Pi_n(\mathbb{R}^k - K) + \Phi_n(\mathbb{R}^k - K)\big). \tag{14.10}$$

As a result of the triangle inequality, we then find that the difference in total-variation distances between Π_n and Φ_n on the one hand and Π_n^K and Φ_n^K on the other is bounded above by the expression on the right in the above display (which is independent of B).

Define A_n, B_n to be the events that $\Pi_n(K_n) > 0$, $\Phi_n(K_n) > 0$ respectively. On $\Xi_n = A_n \cap B_n$, $\Pi_n^{K_n}$ and $\Phi_n^{K_n}$ are well-defined probability measures. Assumption (14.8) guarantees that $P_0^n(\Xi_n)$ converges to 1. Restricting attention to the event Ξ_n in the above upon substitution of the sequence (K_n) and using (14.8) for the limit of (14.10) we find (14.9), where it is understood that the conditional probabilities on the *l.h.s.* are well-defined with probability growing to 1. □

To apply the above lemma in the concluding steps of the proof of theorem 14.3, rate conditions for both posterior and limiting normal sequences are needed. The rate condition (14.3) for the posterior is assumed and the following lemma demonstrates that its analog for the sequence of normals is satisfied when the sequence of centre points Δ_{n,θ_0} is uniformly tight.

Lemma 14.7 *Let K_n be a sequence of balls centred on the origin with radii $M_n \to \infty$. Let (Φ_n) be a sequence of normal distributions (with fixed covariance matrix V) located respectively at the (random) points $(\Delta_n) \subset \mathbb{R}^k$. If the sequence Δ_n is uniformly tight, then:*

$$\Phi_n(\mathbb{R}^k - K_n) = N_{\Delta_n,V}(\mathbb{R}^k - K_n) \xrightarrow{P_0} 0.$$

Proof Let $\delta > 0$ be given. Uniform tightness of the sequence (Δ_n) implies the existence of a constant $L > 0$ such that:

$$\sup_{n\geq 1} P_0^n(\|\Delta_n\| \geq L) \leq \delta.$$

For all $n \geq 1$, call $A_n = \{\|\Delta_n\| \geq L\}$. Let $\mu \in \mathbb{R}^k$ be given. Since $N(\mu, V)$ is tight, for every given $\epsilon > 0$, there exists a constant L' such that $N_{\mu,V}(B(\mu, L')) \geq 1 - \epsilon$ (where $B(\mu, L')$ defines a ball of radius L' around the point μ. Assuming that $\mu \leq L$, $B(\mu, L') \subset B(0, L + L')$ so that with $M = L + L'$, $N_{\mu,V}(B(0, M)) \geq 1 - \epsilon$ for all μ such that $\|\mu\| \leq L$. Choose $N \geq 1$ such that $M_n \geq M$ for all $n \geq N$. Let $n \geq N$ be given. Then:

$$\begin{aligned} P_0^n\bigl(\Phi_n(\mathbb{R}^k - B(0, M_n) > \epsilon\bigr) &\leq P_0^n\bigl(A_n\bigr) + P_0^n\Bigl(\bigl\{\Phi_n(\mathbb{R}^k - B(0, M_n) > \epsilon\bigr\} \cap A_n^c\Bigr) \\ &\leq \delta + P_0^n\Bigl(\bigl\{N_{\Delta_n,V}(B(0, M_n)^c) > \epsilon\bigr\} \cap A_n^c\Bigr) \end{aligned} \tag{14.11}$$

Note that on the complement of A_n, $\|\Delta_n\| < L$, so:

$$N_{\Delta_n,V}(B(0, M_n)^c) \leq 1 - N_{\Delta_n,V}(B(0, M)) \leq 1 - \inf_{\|\mu\|\leq L} N_{\mu,V}(B(0, M)) \leq \epsilon,$$

and we conclude that the last term on the *r.h.s.* of (14.11) equals zero. □

Exercises

14.1 Show that for any probability measure P, Q, there exists a σ-finite measure μ such that $P, Q \ll \mu$. Then prove (14.1).

14.2 Let $\Theta = (0, \infty)$ and $\mathcal{P} = \{N(0, \theta^2) : \theta \in \Theta\}$. Let Π be a thick prior on Θ. Show that this model satisfies the conditions of the Bernstein-von Mises theorem 14.1. Find the problematic range of parameter values in this model. *(Hint: calculate the Fisher information, find a problematic limit for it and describe the effect on the limiting sequence of normal distributions for certain parameter values.)*

14.3 Approximation in measure from within by compact subsets has a deep background in analysis. Central is the notion of a Radon measure. Given a Hausdorff topological space Θ, a *Radon measure* Π is a Borel measure that is *locally finite* (meaning that any $\theta \in \Theta$ has a neighbourhood U such that $\Pi(U) < \infty$) and *inner regular* (meaning that for any subset $S \subset \Theta$ and any $\epsilon > 0$, there exists a compact $K \subset S$ such that $\mu(S - K) < \epsilon$). Show that any probability measure on a Polish space is Radon. *(NB: This statement can be generalized to continuous images of Polish spaces, known as Suslin spaces.)*

15

The semiparametric Bernstein-von Mises theorem

Neither the frequentist theory of asymptotic optimality for regular estimation in smooth models, nor Theorem 14.1 generalize fully to nonparametric estimation problems. Examples of the failure of the Bernstein-Von Mises limit in infinite-dimensional problems (with regard to the *full* parameter) can be found in Freedman (1999). Freedman initiated a discussion concerning the merits of Bayesian methods in nonparametric problems as early as 1963, showing that even with a natural and seemingly innocuous choice of the nonparametric prior, posterior inconsistency may result (Freedman (1963)). This warning against instances of inconsistency due to ill-advised nonparametric priors was reiterated in the literature many times over, for example in Cox (1993) and in Diaconis and Freedman (1986, 1998). However, general conditions for Bayesian consistency were formulated by Schwartz as early as 1965 (Schwartz (1965)) and positive results on posterior rates of convergence in the same spirit were obtained in Ghosal, Ghosh and van der Vaart (2000) (see also, Shen and Wasserman (2001)). The combined message of negative and positive results appears to be that the choice of a nonparametric prior is a sensitive one that leaves room for unintended consequences unless due care is taken.

As we shall see in the following, this lesson must also be taken seriously when one asks the question whether the posterior for the parameter of interest in a semiparametric estimation problem displays limiting behaviour of the type (14.4). But before we formulate and prove a semiparametric version of the Bernstein-Von Mises theorem in this chapter and next, let us have a brief look at the semiparametric theory of efficient point-estimation in differentiable models.

15.1 Efficiency in semiparametric estimation problems

As argued in lecture 1, there are serious limitations to the usefulness of parametric models in statistics: the frequentist assumption that the true distribution of the data belongs to the model, when that model is such a narrowly defined family of distributions, is a very stringent one. Under those conditions, the possibility that the true distribution of the data is *not* in the model is by far the more likely one (in which case we say the the model is *misspecified*). Moreover, verification of this assumption is difficult at best and, more often, wholly impossible. Although subjectivist Bayesians avoid this awkward assumption and therefore use parametric models more liberally, one can question the generalized, universal value of conclusions based on a subjectivist view that is equally narrowly defined.

For that reason, modern statistics studies nonparametric models. To illustrate, given a sample space $\mathscr{X}$ we can choose to take for our model the collection $\mathfrak{M}(\mathscr{X})$ of *all* probability distributions. Then, in the (frequentist) perspective that the data has *some* distribution, the model can never be misspecified. In cases where the data is *i.i.d.*, the empirical measure appears to be a suitable estimator. But more often, the nature of the estimation problem allows one to be more specific concerning the model, or at least, hope to approximate the truth closely enough to guarantee validity of inferential conclusions.

The desire to answer specific statistical questions without concessions that imply gross model misspecification, is what motivates semiparametric statistics: suppose that we are interested in estimation of one (or a finite number of) real-valued aspect(s) of the distribution of the data, like its expectation, variance, or more complicated functionals, like its α-quantile or the L_2-norm of the associated density. One could devise a parametric model for the purpose, but given the above objection of misspecification, one prefers to model the distribution of the data in maximal appropriate generality and estimate aspects of interest based thereon.

15.1.1 Examples of semiparametric problems

Although the more general formulation would concern a nonparametric model $\mathcal{P}$ with a finite-dimensional vector of functionals $\theta: \mathcal{P} \to \mathbb{R}^k$, representing the aspects of interest, we choose to parametrize model distributions in terms of a finite-dimensional *parameter of interst* $\theta \in \Theta$, for an open $\Theta \subset \mathbb{R}^k$, and an infinite-dimensional *nuisance parameter* $\eta \in H$; the non-parametric model is then represented as $\mathcal{P} = \{P_{\theta,\eta}: \theta \in \Theta, \eta \in H\}$. Of course, we impose differentiability (or, more specifically, the LAN property) on the model in a suitable way and we intend to estimate θ efficiently. We mention three popular examples of such estimation problems but note that there are many more.

Example 15.1 (Errors-in-variables regression)
Consider a random vector (X, Y) assumed to follow a regression relation of the form : $Y = \alpha + \beta\, X + e$, for $(\alpha, \beta) \in \mathbb{R}^2$ and an error e that is independent of X and satisfies $Ee = 0$. These widely used models for linear random-design regression suffer from a phenomenon known as *regression dilution* (or *attenuation bias*): although noise in the regressed variable Y is accounted for by the error e, noise in the design points X biases the estimated slope $\hat{\alpha}$ towards zero! (To see why, imagine an exaggerated amount of noise in X which would blur any linear relationship between X and Y beyond recognition and lead to estimates of α close to zero.) For that reason, generalizations of the model have been proposed; most prominent is the semiparametric *errors-in-variables* model, which accounts for noise in the design points explicitly: the model formulates observed (X, Y) and an unobserved random variable Z, related through the regression equations,

$$X = Z + e, \quad \text{and} \quad Y = g_\theta(Z) + f,$$

where the errors (e, f) are assumed independent of Z and such that $Ee = Ef = 0$. Although variations are possible, the most popular formulation of the model involves a family of regression functions that is linear: $g_{\alpha,\beta}(z) = \alpha + \beta\, z$ and a completely unknown distribution F for the unobserved $Z \sim F$. Interest then goes to (efficient) estimation of the parameter $\theta = (\alpha, \beta)$, while treating $\eta = F$ as the nuisance parameter. (For an overview, see Anderson (1985).)

Example 15.2 (Cox' proportional hazards model)

In medical studies (but also many in other disciplines) one is often interested in the relationship between the time of "survival" (which can mean anything from time until actual death, to onset of a symptom, or detection of a certain protein in a patients blood, etc.) and covariates believed to be of influence. Observations consist of pairs (T, Z) associated with individual patients, where T is the survival time and Z is a vector of covariates. The probability of non-survival between t and $t + dt$, given survival up to time t is called the *hazard function* $\lambda(t)$,

$$\lambda(t)\,dt = P\big(\,t \leq T \leq t + dt\;\big|\;T \geq t\,\big).$$

The *Cox proportional hazards model* prescribes that the *conditional hazard function* given Z is of the form,

$$\lambda(t|Z)\,dt = e^{\theta^T Z}\,\lambda_0(t),$$

where λ_0 is the so-called *baseline hazard function*. The interpretation of the parameter of interest θ is easily established: if, for example, the component $Z_i \in \{0, 1\}$ describes presence (or not) of certain characteristics in the patient (*e.g.* $Z = 0$ for a non-smoker and $Z = 1$ for a smoker), then e^{θ_i} is the ratio of hazard rates between two patients that differ only in that one characteristic Z_i. The parameter of interest is the vector of θ, while the baseline hazard rate is treated as an unknown nuisance parameter. (For a discussion of the semiparametric Bernstein-Von Mises theorem in the proportional hazards model, see Kim (2006) and Castillo (2008)).

Example 15.3 (Partial linear regression)
Consider a situation in which one observes a vector $(Y; U, V)$ of random variables, assumed related through the regression equation,

$$Y = \theta\, U + \eta(V) + e,$$

with e independent of the pair (U, V) and such that $Ee = 0$, usually assumed normally distributed. The rationale behind this model would arise from situations where one is observing a linear relationship between two random variables Y and U, contaminated by an additive influence from V of largely unknown form. The parameter $\theta \in \mathbb{R}$ is of interest while the nuisance η is from some infinite-dimensional function space H. *(NB: The distribution P of (U, V) is subject only to certain qualitative restrictions and, as such, forms another nonparametric nuisance component in the model. However, (U, V) is* ancillary*: the P-dependent factor in the likelihood does not vary with (θ, η) and threrefore cancels in the likelihood ratios that control (maximum-likelihood estimates and) posterior distributions.)* Estimation in the partial linear regression model is discussed in lecture 16.

15.1.2 Semiparametric efficiency

Before we consider the problem of semiparametric estimation from a Bayesian perspective, we give a brief account of the central argument regarding strategies for point-estimation of θ. It is assumed that the model $\mathcal{P}$ is dominated by a σ-finite measure with densities $p_{\theta,\eta}$. Furthermore, we assume that the parametrization in terms of θ and η is identifiable and that the true distribution of the data P_0 is contained in the model, implying that there exist unique $\theta_0 \in \Theta$, $\eta_0 \in H$ such that $P_0 = P_{\theta_0,\eta_0}$.

The strategy for finding efficient estimators for θ_0 is based on the following: suppose that $\mathcal{P}_0$ is a submodel of $\mathcal{P}$ and that $\mathcal{P}_0$ contains P_0. Then estimation of θ_0 in the model $\mathcal{P}_0$ is no harder than it is in $\mathcal{P}$. For instance, if one applies this self-evident truth to LAN models (*c.f.* theorem 13.9), one reaches the conclusion that the Fisher information in the larger model is smaller than or equal to that in the smaller model. So, using the same amount of data, estimation of the parameter θ can be done more accurately in the smaller model than in the larger one (in the large sample limit). Such is the price one pays for use of a more general model. Semiparametric information bounds are obtained as infima over the information bounds one obtains from a collection of smooth, finite-dimensional submodels. That collection has to be somehow "rich enough" to capture the true, sharp information bound for (regular) semiparametric estimators. The following is a simplified reflection of the argument put forth in van der Vaart (1998), chapter 25.

Assume that the parameter of interest is one-dimensional and define smooth submodels as follows: for open neighbourhoods U of θ_0, consider maps $\gamma\colon U \to \mathcal{P}\colon \theta \mapsto P_\theta$, such that $P_{\theta=\theta_0} = P_0$ and, for all $\theta \in U$, $P_\theta = P_{\theta,\eta}$ for some $\eta \in H$. Assume that there exists a P_0-square-integrable score function $\dot{\ell}$ such that $P_0\dot{\ell} = 0$ and the LAN property is satisfied:

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0+n^{-1/2}h_n}}{p_0}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h\,\dot{\ell}(X_i) - \tfrac{1}{2}h^2 P_0\dot{\ell}^2 + o_{P_0}(1),$$

for $h_n \to h$. Let $\mathscr{S}$ denote a collection of such smooth submodels. The corresponding collection of score functions $\{\dot{\ell} \in L_2(P_0)\colon \gamma \in \mathscr{S}\}$ may not be closed, but there exists an $\tilde{\ell}_{\mathscr{S}}$ in its $L_2(P_0)$-closure such that:

$$\tilde{I}_{\mathscr{S}} := P_0\tilde{\ell}_{\mathscr{S}}^2 = \inf_{\{\dot{\ell}:\gamma\in\mathscr{S}\}} P_0\dot{\ell}^2. \tag{15.1}$$

With a slight abuse of nomenclature, we refer to $\tilde{\ell}_{\mathscr{S}}$ and $\tilde{I}_{\mathscr{S}}$ as the *efficient score function* and *efficient Fisher information* for θ_0 at P_{θ_0,η_0}, relative to $\mathscr{S}$. The efficient Fisher information $\tilde{I}_{\mathscr{S}}$ captures the notion of an "infimal Fisher information" (over $\mathscr{S}$) alluded to above. Clearly, $\tilde{I}_{\mathscr{S}}$ decreases if we enlarge the collection $\mathscr{S}$.

To arrive at a formulation of efficiency in semiparametric context, let $\mathscr{S}$ denote a collection of LAN submodels and call any estimator sequence (T_n) for θ_0 *regular* with respect to $\mathscr{S}$, if (T_n) is regular as an estimator for θ_0 in all $\gamma \in \mathscr{S}$ (*c.f.* definition 13.7). Theorem 13.9 applies in any $\gamma \in \mathscr{S}$ so we obtain a collection of Fisher information bounds, one for each $\gamma \in \mathscr{S}$. This implies that for any (T_n) regular with respect to $\mathscr{S}$, a convolution theorem can be formulated in which the inverse *efficient* Fisher information $\tilde{I}_{\mathscr{S}}$ represents a lower bound to estimation accuracy. For the following theorem, which can be found as theorem 25.20 in van der Vaart (1998), define the *tangent set* to be $\{a\dot{\ell}\colon a \in [0,\infty), \gamma \in \mathscr{S}\} \subset L_2(P_0)$.

Theorem 15.4 *(Semiparametric convolution theorem)*
Let $\Theta \subset \mathbb{R}^k$ be open; let H be an infinite-dimensional nuisance space and let $\mathcal{P}$ be the corresponding semiparametric model. Let a collection of smooth submodels $\mathscr{S}$ be given. Assume that the true distribution of the i.i.d. data is P_{θ_0,η_0}. For any estimator sequence (T_n) that is regular with respect to $\mathscr{S}$, the asymptotic covariance matrix is lower bounded by $\tilde{I}_{\mathscr{S}}$. Furthermore, if the tangent set is a convex cone, the limit distribution of (T_n) is of the form $N(0,\tilde{I}_{\mathscr{S}}^{-1}) \star M$ for some probability distribution M.

As noted from the start of this section, if the collection $\mathscr{S}$ of smooth submodels is somehow not "rich enough", the resultant "efficient" Fisher information may not be truely infimal and hence, may not give rise to a *sharp* bound on the asymptotic variance of regular estimators. As a result optimal regular sequences (T_n) in theorem 15.4 (in the sense that $M = \delta_0$) may not exist. Of course, there is the option of maximizing $\mathscr{S}$ to contain *all* LAN submodels containing P_0.

Definition 15.5 The efficient score function and efficient Fisher information relative to the *maximal* $\mathscr{S}$ containing all LAN submodels are referred to as *the* efficient score function and *the* efficient Fisher information, denoted by $\tilde{\ell}_{\theta_0,\eta_0}$ and $\tilde{I}_{\theta_0,\eta_0}$ respectively.

But the maximal collection of LAN submodels is hard to handle in any practical sense because its definition is implicit. Usually the way one proceeds in any given model is by proving the LAN property for some $\mathscr{S}$ and defining a clever proposal for a regular point-estimator, with the goal of demonstrating that its limit distribution *attains* the lower bound implied by the semiparametric convolution theorem. (Compare this with the manner in which we concluded that, under the conditions of theorem 13.3, the parametric ML estimator is efficient.)

Theorem 15.6 *Let $\mathscr{S}$ be a collection of smooth submodels of $\mathcal{P}$ with corresponding efficient Fisher information $\tilde{I}_{\mathscr{S}}$. Let (T_n) be a regular estimator sequence for the parameter of interest. If (T_n) is asymptotically normal with asymptotic covariance $\tilde{I}_{\mathscr{S}}^{-1}$, then $\tilde{I}_{\theta_0,\eta_0} = \tilde{I}_{\mathscr{S}}$ and (T_n) is best-regular.*

Compare semiparametric and parametric estimations problems as follows: suppose we would know the true value η_0 of the nuisance parameter: in that case, we could estimate θ_0 within the parametric model $\gamma_0 = \{P_{\theta,\eta_0} : \theta \in \Theta\}$, with corresponding score $\dot{\ell}_{\theta_0,\eta_0}$ and Fisher information I_{θ_0,η_0}. This I_{θ_0,η_0} provides an information bound that is greater than, or equal to $\tilde{I}_{\theta_0,\eta_0}$ since γ_0 is an element of the maximal $\mathscr{S}$. The deterioration in asymptotic estimation accuracy implied by the transition from I_{θ_0,η_0} to $\tilde{I}_{\theta_0,\eta_0}$ reflects the use of a (far) more general model for the data through inclusion of a non-parametric nuisance.

Like in the parametric theory of efficient estimation, the convolution theorem gives rise to the notion of a optimal regular semiparametric estimator sequence: if (T_n) is regular and attains efficiency, *i.e.* if,

$$n^{1/2}(T_n - \theta_0) \overset{\theta_0,\eta_0}{\rightsquigarrow} N(0, \tilde{I}_{\theta_0,\eta_0}^{-1}),$$

then (T_n) is said to be *best-regular*. Semiparametric estimators (T_n) for θ_0 are best-regular *if and only if* the (T_n) are asymptotically linear, that is,

$$n^{1/2}(T_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}_{\theta_0,\eta_0}^{-1} \tilde{\ell}_{\theta_0,\eta_0}(X_i) + o_{P_\theta}(1), \tag{15.2}$$

(For a proof, see lemma 25.23 in van der Vaart (1998)).

So, one reasons, in order to prove semiparametric efficiency, it would be enough to find a single smooth submodel $\tilde{\gamma}$ for which the Fisher information equals the efficient Fisher information. Any estimator sequence that is regular is then best-regular for θ_0 in $\mathcal{P}$. If it exists, such a submodel is called a *least-favourable submodel*. Somewhat disappointingly, in many semiparametric problems, least-favourable submodels do not exist. That eventuality does not impede the definition of the efficient score, because that is an element of the *closure* of the tangent set. But it does limit the applicability of constructions that depend on the existence of least-favourable submodels, such as the Bernstein-Von Mises theorem presented later in this lecture.

15.2 Bayesian semiparametric statistics

Building on the analogy drawn between efficiency in smooth, parametric models and the present, semiparametric efficiency question, we look for general sufficient conditions on model and prior such that the *marginal posterior for the parameter of interest* satisfies,

$$\sup_B \Bigl| \Pi\bigl(\sqrt{n}(\theta - \theta_0) \in B \bigm| X_1, \ldots, X_n \bigr) - N_{\tilde{\Delta}_n, \tilde{I}^{-1}_{\theta_0,\eta_0}}(B) \Bigr| \to 0, \tag{15.3}$$

in P_{θ_0}-probability, where,

$$\tilde{\Delta}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{I}^{-1}_{\theta_0,\eta_0} \tilde{\ell}_{\theta_0,\eta_0}(X_i). \tag{15.4}$$

Here $\tilde{\ell}_{\theta,\eta}$ denotes the efficient score function and $\tilde{I}_{\theta,\eta}$ the efficient Fisher information (assumed to be non-singular at (θ_0, η_0)). Assertion (15.3) often implies efficiency of point-estimators like the posterior median, mode or mean and always leads to asymptotic identification of credible regions with efficient confidence regions. Like before, if C is a credible set in Θ, (15.3) guarantees that posterior coverage and coverage under the limiting normal for C are (close to) equal. Because the limiting normals correspond to the asymptotic sampling distributions for efficient point-estimators, (15.3) enables interpretation of credible sets as asymptotically efficient confidence regions. From a practical point of view, the latter conclusion has an important implication: whereas it can be hard to compute optimal semiparametric confidence regions directly (not least of all because one has to estimate the efficient Fisher information), simulation of a large sample from the marginal posterior (*e.g.* by MCMC techniques, see Robert (2001)) is sometimes comparatively straightforward.

Instances of the Bernstein-Von Mises limit have been studied in various semiparametric models: several papers have provided studies of asymptotic normality of posterior distributions for models from survival analysis. Particularly, Kim and Lee (2004) show that the *infinite-dimensional* posterior for the cumulative hazard function under right-censoring converges at rate $n^{-1/2}$ to a Gaussian centred at the Aalen-Nelson estimator for a class of neutral-to-the-right process priors. In Kim (2006), the posterior for the baseline cumulative hazard function and regression coefficients in Cox' proportional hazard model are considered with similar priors. Castillo (2008) considers marginal posteriors in Cox' proportional hazards model and Stein's symmetric location problem from a unified point of view. A general approach has been given in Shen (2002) but his conditions may prove somewhat hard to verify in examples. Cheng and Kosorok (2008) give a general perspective too, proving weak convergence of the posterior under sufficient conditions. Rivoirard and Rousseau (2009) prove a version for linear functionals over the model, using a class of nonparametric priors based on infinite-dimensional exponential families. Boucheron and Gassiat (2009) consider the Bernstein-Von Mises theorem for families of discrete distributions. Johnstone (2010) studies various marginal posteriors in the Gaussian sequence model.

15.3 The semiparametric Bernstein-Von Mises theorem

Consider estimation of a functional $\theta\colon \mathcal{P} \to \mathbb{R}^k$ on a dominated nonparametric model $\mathcal{P}$ with metric g, based on a sample $X_1, X_2, \ldots$, distributed *i.i.d.* according to $P_0 \in \mathcal{P}$. We introduce a prior Π on $\mathcal{P}$ and consider the subsequent sequence of posteriors,

$$\Pi\bigl(A \bigm| X_1, \ldots, X_n \bigr) = \int_A \prod_{i=1}^n p(X_i)\, d\Pi(P) \Bigm/ \int_{\mathcal{P}} \prod_{i=1}^n p(X_i)\, d\Pi(P), \tag{15.5}$$

where A is any measurable model subset. Typically, optimal (*e.g.* minimax) nonparametric posterior rates of convergence (see Ghosal *et al.* (2000)) are powers of n (possibly modified by a slowly varying function) that converge to zero more slowly than the parametric $n^{-1/2}$-rate. Estimators for θ may be derived by 'plugging in' a nonparametric estimate, *c.f.* $\hat{\theta} = \theta(\hat{P})$, but optimality in rate or asymptotic variance cannot be expected to obtain generically in this way. This does not preclude efficient estimation of real-valued aspects of P_0: parametrize the model in terms of a finite-dimensional *parameter of interest* $\theta \in \Theta$ and a *nuisance parameter* $\eta \in H$ where Θ is open in $\mathbb{R}^k$ and (H, d_H) an infinite-dimensional metric space: $\mathcal{P} = \{\, P_{\theta,\eta} : \theta \in \Theta, \eta \in H \,\}$. Assuming identifiability, there exist unique $\theta_0 \in \Theta$, $\eta_0 \in H$ such that $P_0 = P_{\theta_0,\eta_0}$. Assuming measurability of the map $(\theta, \eta) \mapsto P_{\theta,\eta}$, we place a product prior $\Pi_\Theta \times \Pi_H$ on $\Theta \times H$ to define a prior on $\mathcal{P}$. Parametric rates for the marginal posterior of θ are achievable because it is possible for contraction of the full posterior to occur anisotropically, that is, at rate $n^{-1/2}$ along the θ-direction, but at a slower, nonparametric rate (ρ_n) along the η-directions.

15.3.1 Steps in the proof

The proof of (15.3) will consist of three steps: in section 17.1, we show that the posterior concentrates its mass around so-called *least-favourable submodels*. In the second step (see section 17.2), we show that this implies local asymptotic normality for integrals of the likelihood over H, with the efficient score determining the expansion. In section 17.3, it is shown that these LAN integrals induce asymptotic normality of the marginal posterior, analogous to the way local asymptotic normality of parametric likelihoods induces the parametric Bernstein-Von Mises theorem.

To see why asymptotic accumulation of posterior mass occurs around so-called least-favourable submodels, a crude argument departs from the observation that, according to (15.5), posterior concentration occurs in regions of the model with relatively high likelihood (barring inhomogeneities of the prior). Asymptotically, such regions are characterized by close-to-minimal Kullback-Leibler divergence with respect to P_0. To exploit this, let us assume that for each θ in a neighbourhood U_0 of θ_0, there exists a unique minimizer $\eta^*(\theta)$ of the Kullback-Leibler divergence,

$$-P_0 \log \frac{p_{\theta,\eta^*(\theta)}}{p_{\theta_0,\eta_0}} = \inf_{\eta \in H} \Bigl(-P_0 \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}} \Bigr), \tag{15.6}$$

giving rise to a submodel $\mathcal{P}^* = \{P^*_\theta = P_{\theta,\eta^*(\theta)} : \theta \in U_0\}$. As is well-known (Severini (1992)), if $\mathcal{P}^*$ is smooth it constitutes a least-favourable submodel and scores along $\mathcal{P}^*$ are efficient. (In subsequent sections it is not required that $\mathcal{P}^*$ is defined by (15.6), only that $\mathcal{P}^*$ is least-favourable.) Neighbourhoods of $\mathcal{P}^*$ are described with Hellinger balls in H of radius $\rho > 0$ around $\eta^*(\theta)$, for all $\theta \in U_0$:

$$D(\theta, \rho) = \{\, \eta \in H : d_H(\eta, \eta^*(\theta)) < \rho \,\}. \tag{15.7}$$

To give a more precise argument for posterior concentration around $\eta^*(\theta)$, consider the posterior for η, *given* $\theta \in U_0$; unless θ happens to be equal to θ_0, the submodel $\mathcal{P}_\theta = \{P_{\theta,\eta} : \eta \in H\}$ is misspecified. Kleijn and van der Vaart (2006) show that the misspecified posterior concentrates asymptotically in any (Hellinger) neighbourhood of the point of minimal Kullback-Leibler divergence with respect to the true distribution of the data. Applied to $\mathcal{P}_\theta$, we see that $D(\theta, \rho)$ receives asymptotic posterior probability one for any $\rho > 0$. For posterior concentration to occur sufficient prior mass must be present in certain Kullback-Leibler-type neighbourhoods. In the present contex, these neighbourhoods can be defined as:

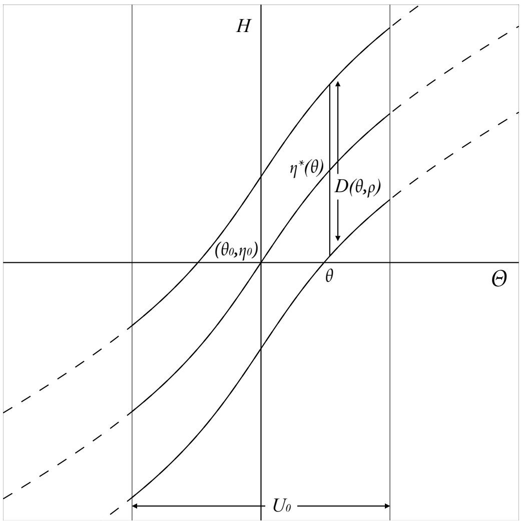

Figure 15.1 A neighbourhood of (θ_0, η_0). Shown are the least-favourable curve $\{(\theta, \eta^*(\theta)) : \theta \in U_0\}$ and (for fixed θ and $\rho > 0$) the neighbourhood $D(\theta, \rho)$ of $\eta^*(\theta)$. The sets $D(\theta, \rho)$ are expected to capture (θ-conditional) posterior mass one asymptotically, for all $\rho > 0$ and $\theta \in U_0$.

$$\begin{aligned} K_n(\rho, M) = \Biggl\{ \eta \in H : P_0 \Bigl(\sup_{\|h\| \le M} -\log \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta_0}} \Bigr) \le \rho^2, \\ P_0 \Bigl(\sup_{\|h\| \le M} -\log \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta_0}} \Bigr)^2 \le \rho^2 \Biggr\}, \end{aligned} \tag{15.8}$$

for $\rho > 0$ and $M > 0$. If this type of posterior convergence occurs with an appropriate form of uniformity over the relevant values of θ (see ‘consistency under perturbation’, section 17.1), one expects that the nonparametric posterior contracts into Hellinger neighbourhoods of the curve $\theta \mapsto (\theta, \eta^*(\theta))$ (theorem 17.1 and corollary 17.3).

To introduce the second step, consider (15.5) with $A = B \times H$ for some measurable $B \subset \Theta$. Since the prior is of product form, $\Pi = \Pi_\Theta \times \Pi_H$, the marginal posterior for the parameter $\theta \in \Theta$ depends on the nuisance factor only through the integrated likelihood ratio,

$$S_n \colon \Theta \to \mathbb{R} \colon \theta \mapsto \int_H \prod_{i=1}^n \frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta), \tag{15.9}$$

where we have introduced factors $p_{\theta_0,\eta_0}(X_i)$ in the denominator for later convenience, see (17.28). (The localized version of (15.9) is denoted $h \mapsto s_n(h)$, see (17.15).) The map S_n is to be viewed in a role similar to that of the *profile likelihood* in semiparametric maximum-likelihood methods (see, *e.g.*, Severini and Wong (1992) and Murphy and van der Vaart (2000)), in the sense that S_n embodies the intermediate stage between nonparametric and semiparametric steps of the estimation procedure.

We impose smoothness through stochastic local asymptotic normality, *c.f.* definition 14.2, on least-favourable submodels. Although formally only a convenience, the presentation benefits from an *adaptive* reparametrization (see section 2.4 of Bickel, Klaassen, Ritov, Wellner (1998)): based on the least-favourable submodel η^*, we define for all $\theta \in U_0$, $\eta \in H$:

$$(\theta, \eta(\theta,\zeta)) = (\theta, \eta^*(\theta) + \zeta), \quad (\theta, \zeta(\theta,\eta)) = (\theta, \eta - \eta^*(\theta)), \tag{15.10}$$

and we introduce the notation $Q_{\theta,\zeta} = P_{\theta,\eta(\theta,\zeta)}$. With $\zeta = 0$, $\theta \mapsto Q_{\theta,0}$ describes the least-favourable submodel $\mathcal{P}^*$ and with a non-zero value of ζ, $\theta \mapsto Q_{\theta,\zeta}$ describes a version thereof, translated over a nuisance direction (see figure 15.2). Expressed in terms of the metric $r_H(\zeta_1,\zeta_2) = H(Q_{\theta_0,\zeta_1}, Q_{\theta_0,\zeta_2})$, the sets $D(\theta,\rho)$ are mapped to open balls $B(\rho) = \{\zeta \in H \colon r_H(\zeta, 0) < \rho\}$ centred at the origin $\zeta = 0$,

$$\{P_{\theta,\eta} \colon \theta \in U_0, \eta \in D(\theta,\rho)\} = \{Q_{\theta,\zeta} \colon \theta \in U_0, \zeta \in B(\rho)\}.$$

In the formulation of theorem 15.7 there is a domination condition based on the quantities,

$$U_n(\rho, h) = \sup_{\zeta \in B(\rho)} Q^n_{\theta_0,\zeta} \left(\prod_{i=1}^n \frac{q_{\theta_n(h),\zeta}}{q_{\theta_0,\zeta}}(X_i) \right),$$

for all $\rho > 0$ and $h \in \mathbb{R}^k$. Below, it is required that there exists a sequence (ρ_n) with $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$, such that, for every *bounded*, stochastic sequence (h_n), $U(\rho_n, h_n) = O(1)$ (where the expectation concerns the stochastic dependence of h_n as well, see *Complements* at the end of this lecture). For a single, fixed ζ, the requirement says that the likelihood ratio remains integrable when we replace $\theta_n(h_n)$ by the maximum-likelihood estimator $\hat\theta_n(X_1, \ldots, X_n)$. Lemma 17.7 demonstrates that ordinary differentiability of the likelihood-ratio with respect to h, combined with a uniform upper bound on certain Fisher information coefficients suffices to satisfy $U(\rho_n, h_n) = O(1)$ for all bounded, stochastic (h_n) and every $\rho_n \downarrow 0$.

The second step of the proof can now be summarized as follows: assuming stochastic

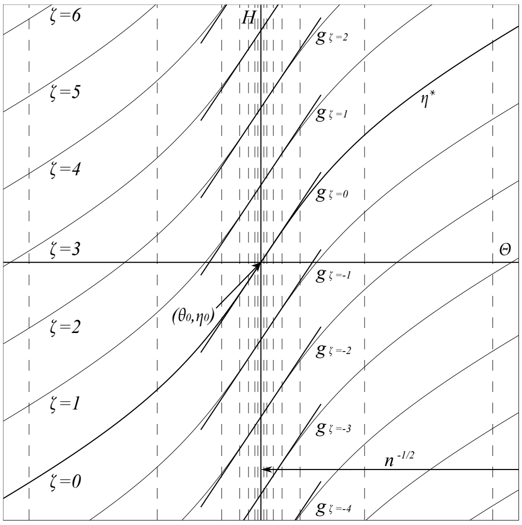

Figure 15.2 A neighbourhood of (θ_0, η_0). Curved lines represent sets $\{(\theta, \zeta) \colon \theta \in U_0\}$ for fixed ζ. The curve through $\zeta = 0$ parametrizes the least-favourable submodel. Vertical dashed lines delimit regions such that $\|\theta - \theta_0\| \leq n^{-1/2}$. Also indicated are directions along which the likelihood is expanded, with score functions g_ζ.

LAN of the model, contraction of the nuisance posterior as in figure 15.1 and said domination condition are enough to turn LAN expansions for the integrand in (15.9) into a single LAN expansion for S_n. The latter is determined by the efficient score, because the locus of posterior concentration, $\mathcal{P}^*$, is a least-favourable submodel (see theorem 17.6).

The third step is based on two obervations: firstly, in a semiparametric problem the integrals S_n appear in the expression for the marginal posterior in exactly the same way as parametric likelihood ratios appear in the posterior for parametric problems. Secondly, the parametric Bernstein-Von Mises proof depends on likelihood ratios *only* through the LAN property. As a consequence, local asymptotic normality for S_n offers the possibility to apply Le Cam's proof of posterior asymptotic normality in semiparametric context. If, in addition we impose contraction at parametric rate for the marginal posterior, the LAN expansion of S_n leads to the conclusion that the marginal posterior satisfies the Bernstein-Von Mises assertion (15.3) (see theorem 17.8).

15.3.2 Formulation of the theorem

Before we state the main result of this lecture, general conditions imposed on models and priors are formulated.

(i) *Model assumptions*
Throughout the remainder, $\mathcal{P}$ is assumed to be well-specified and dominated by a σ-finite measure on the samplespace and parametrized identifiably on $\Theta \times H$, with $\Theta \subset \mathbb{R}^k$ open and H a subset of a metric vector-space with metric d_H. Smoothness of the model

is required but mentioned explicitly throughout. We also assume that there exists an open neighbourhood $U_0 \subset \Theta$ of θ_0 on which a least-favourable submodel $\eta^*\colon U_0 \to H$ is defined.

(ii) *Prior assumptions*
With regard to the prior Π we follow the product structure of the parametrization of $\mathcal{P}$, by endowing the parameterspace $\Theta \times H$ with a product-prior $\Pi_\Theta \times \Pi_H$ defined on a σ-field that includes the Borel σ-field generated by the product-topology. Also, it is assumed that the prior Π_Θ is thick at θ_0.

With the above general considerations for model and prior in mind, we formulate the main theorem.

Theorem 15.7 (Semiparametric Bernstein-Von Mises)
Let $X_1, X_2, \ldots$ be distributed i.i.d.-P_0, with $P_0 \in \mathcal{P}$ and let Π_Θ be thick at θ_0. Suppose that for large enough n, the map $h \mapsto s_n(h)$ is continuous P_0^n-almost-surely. Also assume that $\theta \mapsto Q_{\theta,\zeta}$ is stochastically LAN in the θ-direction, for all ζ in an r_H-neighbourhood of $\zeta = 0$ and that the efficient Fisher information $\tilde{I}_{\theta_0 . \eta_0}$ is non-singular. Furthermore, assume that there exists a sequence (ρ_n) with $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$ such that:

(i) For all $M > 0$, there exists a $K > 0$ such that, for large enough n,

$$\Pi_H\bigl(K_n(\rho_n, M)\bigr) \geq e^{-Kn\rho_n^2}.$$

(ii) For all n large enough, the Hellinger metric entropy satisfies,

$$N\bigl(\rho_n, H, d_H\bigr) \leq e^{n\rho_n^2},$$

and, for every bounded, stochastic (h_n),

(iii) The model satisfies the domination condition,

$$U_n(\rho_n, h_n) = O(1). \tag{15.11}$$

(iv) For all $L > 0$, Hellinger distances satisfy the uniform bound,

$$\sup_{\{\eta \in H : d_H(\eta,\eta_0) \geq L\rho_n\}} \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} = o(1).$$

Finally, suppose that,

(v) For every (M_n), $M_n \to \infty$, the posterior satisfies,

$$\Pi_n\bigl(\, \|h\| \leq M_n \bigm| X_1, \ldots, X_n \,\bigr) \xrightarrow{P_0} 1.$$

Then the sequence of marginal posteriors for θ converges in total variation to a normal distribution,

$$\sup_A \Bigl| \Pi_n\bigl(\, h \in A \bigm| X_1, \ldots, X_n \,\bigr) - N_{\tilde{\Delta}_n, \tilde{I}^{-1}_{\theta_0,\eta_0}}(A) \Bigr| \xrightarrow{P_0} 0, \tag{15.12}$$

centred on $\tilde{\Delta}_n$ with covariance matrix $\tilde{I}^{-1}_{\theta_0,\eta_0}$.

Proof The assertion follows from combination of theorem 17.1, corollary 17.3, theorem 17.6 and theorem 17.8. □

Let us briefly discuss some aspects of the conditions of theorem 15.7. First, consider the required existence of a least-favourable submodel in $\mathscr{P}$. As said, for many semiparametric problems the efficient score function is *not* a proper score in the sense that it corresponds to a smooth submodel. However, there exist sequences of so-called *approximately least-favourable* submodels whose scores converge to the efficient score in L_2. Using such approximations of $\mathscr{P}^*$, our proof will entail extra conditions, but there is no reason to expect problems of an overly restrictive nature. It may therefore be hoped that the result remains largely unchanged, if we turn (15.10) into a sequence of reparametrizations based on suitably chosen approximately least-favourable submodels.

Second consider the rate (ρ_n), which must be slow enough to satisfy condition *(iv)* and is fixed at (or above) the minimax Hellinger rate for estimation of the nuisance with known θ_0 by condition *(ii)*, while satisfying *(i)* and *(iii)* as well. Conditions *(i)* and *(ii)* also arise when considering Hellinger rates for nonparametric posterior convergence and the methods of Ghosal *et al.* (2000) can be applied in the present context with minor modifications. In addition, lemma 17.7 shows that in a wide class of semiparametric models, condition *(iii)* is satisfied for *any* rate sequence (ρ_n). Typically, the numerator in condition *(iv)* is of order $O(n^{-1/2})$, so that condition *(iv)* holds true for any ρ_n such that $n\rho_n^2 \to \infty$. The above enables a rate-free version of the semiparametric Bernstein-Von Mises theorem (corollary 16.1), in which conditions *(i)* and *(ii)* above are weakened to become comparable to those of Schwartz (1965) for nonparametric posterior consistency. Applicability of corollary 16.1 is demonstrated in lecture 16, where the linear coefficient in the partial linear regression model is estimated.

Third, consider condition *(v)* of theorem 15.7: though it is necessary (as it follows from (15.12)), it is hard to formulate straightforward sufficient conditions to satisfy *(v)* in generality. Moreover, condition *(v)* involves the nuisance prior and, as such, imposes another condition on Π_H besides *(i)*. To lessen its influence on Π_H, constructions in section 15.4 either work for all nuisance priors (see lemma 15.8), or require only consistency of the nuisance posterior (see theorem 15.9). The latter is based on the limiting behaviour of posteriors in misspecified parametric models Kleijn (2003), Kleijn and van der Vaart (2007) and allows for the tentative but general observation that a bias (*c.f.* (15.18)) may ruin $n^{-1/2}$-consistency of the marginal posterior, especially if the rate (ρ_n) is sub-optimal. In the example of lecture 16, the 'hard work' stems from condition *(v)* of theorem 15.7: $\alpha > 1/2$ Hölder smoothness and boundedness of the family of regression functions in corollary 16.3 are imposed in order to satisfy this condition. Since conditions *(i)* and *(ii)* appear quite reasonable and conditions *(iii)* and *(iv)* are satisfied relatively easily, condition *(v)* should be viewed as the most complicated in an essential way. For that reason, it forms the next subject in this lecture and we postpone discussion of the proof until after.

15.4 Marginal posterior convergence at parametric rate

Condition (17.29) in theorem 17.8 requires that the posterior measures of a sequence of model subsets of the form,

$$\Theta_n \times H = \big\{(\theta,\eta) \in \Theta \times H : \sqrt{n}\|\theta - \theta_0\| \le M_n\big\}, \tag{15.13}$$

converge to one in P_0-probability, for every sequence (M_n) such that $M_n \to \infty$. Essentially, this condition enables us to restrict the proof of theorem 17.8 to the shrinking domain in which (17.16) applies. In this section, we consider two distinct approaches: the first (lemma 15.8) is based on bounded likelihood ratios (see also condition (B3) of theorem 8.2 in Lehmann and Casella (1998)). The second is based on the behaviour of misspecified parametric posteriors (theorem 15.9). The latter construction illustrates the intricacy of this section's subject most clearly and provides some general insight. Methods proposed here are neither compelling nor exhaustive, we simply put forth several possible approaches and demonstrate the usefulness of one of them in lecture 16.

Lemma 15.8 (Marginal parametric rate (I))
Let the sequence of maps $\theta \mapsto S_n(\theta)$ be P_0-almost-surely continuous and such that (17.16) is satisfied. Furthermore, assume that there exists a constant $C > 0$ such that for any (M_n), $M_n \to \infty$,

$$P_0^n\biggl(\sup_{\eta\in H}\sup_{\theta\in\Theta_n^c}\mathbb{P}_n\log\frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} \le -\frac{C\,M_n^2}{n}\biggr) \to 1. \tag{15.14}$$

Then, for any nuisance prior Π_H and parametric prior Π_Θ, thick at θ_0,

$$\Pi\bigl(\,n^{1/2}\|\theta-\theta_0\| > M_n \;\bigm|\; X_1,\ldots,X_n\,\bigr) \xrightarrow{P_0} 0, \tag{15.15}$$

for any (M_n), $M_n \to \infty$.

Proof Let (M_n), $M_n \to \infty$ be given. Define (A_n) to be the events in (15.14) so that $P_0^n(A_n^c) = o(1)$ by assumption. In addition, let,

$$B_n = \biggl\{\int_\Theta S_n(\theta)\,d\Pi_\Theta(\theta) \ge e^{-\frac{1}{2}C\,M_n^2}\,S_n(\theta_0)\biggr\}.$$

By (17.16) and lemma 15.10, $P_0^n(B_n^c) = o(1)$ as well. Then,

$$\begin{aligned}
P_0^n\Pi(\theta\in\Theta_n^c|X_1,\ldots,X_n) &\le P_0^n\Pi(\theta\in\Theta_n^c|X_1,\ldots,X_n)\,1_{A_n\cap B_n} + o(1)\\
&\le e^{\frac{1}{2}C\,M_n^2}\,P_0^n\biggl(S_n(\theta_0)^{-1}\\
&\quad\times\int_H\int_{\Theta_n^c}\prod_{i=1}^n\frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i)\prod_{i=1}^n\frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}(X_i)\,d\Pi_\Theta\,d\Pi_H\,1_{A_n}\biggr) + o(1) = o(1),
\end{aligned}$$

which proves (15.15). □

Although applicable directly in the model of lecture 16, most other examples would require variations. Particularly, if the full, non-parametric posterior is known to concentrate on a sequence of model subsets (V_n), then lemma 15.8 can be preceded by a decomposition of $\Theta\times H$ over V_n and V_n^c, reducing condition (15.14) to a supremum over V_n^c (see section 2.4 in Kleijn (2003) and the discussion following the following theorem).

Our second approach assumes such concentration of the posterior on model subsets, *e.g.* deriving from non-parametric consistency in a suitable form. Though the proof of theorem 15.9 is rather straightforward, combination with results in misspecified parametric models (Kleijn and van der Vaart (2007)) leads to the observation that marginal parametric rates of convergence can be ruined by a bias.

Theorem 15.9 (Marginal parametric rate (II))
Let Π_Θ and Π_H be given. Assume that there exists a sequence (H_n) of subsets of H, such that the following two conditions hold:

(i) The nuisance posterior concentrates on H_n asymptotically,

$$\Pi\big(\,\eta \in H - H_n \bigm| X_1, \ldots, X_n\big) \xrightarrow{P_0} 0. \tag{15.16}$$

(ii) For every (M_n), $M_n \to \infty$,

$$P_0^n \sup_{\eta \in H_n} \Pi\big(\,n^{1/2}\|\theta - \theta_0\| > M_n \bigm| \eta, X_1, \ldots, X_n\,\big) \to 0. \tag{15.17}$$

Then the marginal posterior for θ concentrates at parametric rate, i.e.,

$$\Pi\big(\,n^{1/2}\|\theta - \theta_0\| > M_n \bigm| \eta, X_1, \ldots, X_n\big) \xrightarrow{P_0} 0,$$

for every sequence (M_n), $M_n \to \infty$,

Proof Let (M_n), $M_n \to \infty$ be given and consider the posterior for the complement of (15.13). By assumption *(i)* of the theorem and Fubini's theorem,

$$\begin{aligned} &P_0^n \Pi\big(\,\theta \in \Theta_n^c \bigm| X_1, \ldots, X_n\big) \\ &\quad\le P_0^n \int_{H_n} \Pi\big(\,\theta \in \Theta_n^c \bigm| \eta, X_1, \ldots, X_n\big)\, d\Pi\big(\,\eta \bigm| X_1, \ldots, X_n\big) + o(1) \\ &\quad\le P_0^n \sup_{\eta \in H_n} \Pi\big(\,n^{1/2}\|\theta - \theta_0\| > M_n \bigm| \eta, X_1, \ldots, X_n\big) + o(1), \end{aligned}$$

the first term of which is $o(1)$ by assumption *(ii)* of the theorem. □

Condition *(ii)* of theorem 15.9 has an interpretation in terms of misspecified parametric models (Kleijn and van der Vaart (2007) and Kleijn (2003)). For fixed $\eta \in H$, the η-conditioned posterior on the parametric model $\mathcal{P}_\eta = \{P_{\theta,\eta} : \theta \in \Theta\}$ is required to concentrate in $n^{-1/2}$-neighbourhoods of θ_0 under P_0. However, this misspecified posterior concentrates around $\Theta^*(\eta) \subset \Theta$, the set of points in Θ where the Kullback-Leibler divergence of $P_{\theta,\eta}$ with respect to P_0 is minimal. Assuming that $\Theta^*(\eta)$ consists of a unique minimizer $\theta^*(\eta)$, the dependence of the Kullback-Leibler divergence on η must be such that,

$$\sup_{\eta \in H_n} \|\theta^*(\eta) - \theta_0\| = o\big(n^{-1/2}\big). \tag{15.18}$$

in order for posterior concentration to occur on the strips (15.13). In other words, minimal Kullback-Leibler divergence may bias the (points of convergence of) η-conditioned parametric posteriors to such an extent that consistency of the marginal posterior for θ is ruined.

The occurrence of this bias is a property of the semiparametric model rather than a perculiarity of the Bayesian approach: when (point-)estimating with solutions to score equations for example, the same bias occurs (see *e.g.* theorem 25.59 in van der Vaart (1998) and subsequent discussion). Frequentist literature also offers some guidance towards mitigation of this circumstance. First of all, it is noted that the bias indicates the existence of a better (*i.e.* bias-less) choice of parametrization to ask the relevant semiparametric question. If the parametrization is fixed, alternative point-estimation methods may resolve bias, for example through replacement of score equations by general estimating equations (see, for example, section 25.9 in van der Vaart (1998)), loosely equivalent to introducing a suitable penalty in a likelihood maximization procedure.

For a so-called *curve-alignment model* with Gaussian prior, the no-bias problem has been addressed and resolved in a fully Bayesian manner by Castillo (2011): like a penalty in an ML procedure, Castillo's (rather subtle choice of) prior guides the procedure away from the biased directions and produces Bernstein-Von Mises efficiency of the marginal posterior. A most interesting question concerns generalization of Castillo's intricate construction to more general Bayesian context.

Referring to definitions (15.9) and (17.15), we conclude this section with a lemma used in the proof of lemma 15.8 to lower-bound the denominator of the marginal posterior.

Lemma 15.10 *Let the sequence of maps $\theta \mapsto S_n(\theta)$ be P_0-almost-surely continuous and such that (17.16) is satisfied. Assume that Π_Θ is thick at θ_0 and denoted by Π_n in the local parametrization in terms of h. Then,*

$$P_0^n\Bigl(\int s_n(h)\,d\Pi_n(h) < a_n\,s_n(0)\Bigr) \to 0, \tag{15.19}$$

for every sequence (a_n), $a_n \downarrow 0$.

Proof Let $M > 0$ be given and define $C = \{h\colon \|h\| \le M\}$. Denote the rest-term in (17.16) by $h \mapsto R_n(h)$. By continuity of $\theta \mapsto S_n(\theta)$, $\sup_{h\in C}|R_n(h)|$ converges to zero in P_0-probability. If we choose a sequence (κ_n) that converges to zero slowly enough, the corresponding events $B_n = \bigl\{\sup_C\bigl|R_n(h)\bigr| \le \kappa_n\bigr\}$, satisfy $P_0^n(B_n) \to 1$. Next, let (K_n), $K_n \to \infty$ be given. There exists a $\pi > 0$ such that $\inf_{h\in C} d\Pi_n/d\mu(h) \ge \pi$, for large enough n. Combining, we find,

$$\begin{aligned} P_0^n\Bigl(\int \frac{s_n(h)}{s_n(0)}\,d\Pi_n(h) \le e^{-K_n^2}\Bigr)& \\ \le P_0^n\Bigl(\Bigl\{\int_C \frac{s_n(h)}{s_n(0)}\,d\mu(h) \le \pi^{-1}\,e^{-K_n^2}\Bigr\} \cap B_n\Bigr) &+ o(1). \end{aligned} \tag{15.20}$$

On B_n, the integral LAN expansion is lower bounded so that, for large enough n,

$$\begin{aligned} P_0^n\Bigl(\Bigl\{\int_C \frac{s_n(h)}{s_n(0)}\,d\mu(h) \le \pi^{-1}\,e^{-K_n^2}\Bigr\} \cap B_n\Bigr)& \\ \le P_0^n\Bigl(\int_C e^{h^T\mathbb{G}_n\tilde\ell_{\theta_0,\eta_0}}\,d\mu(h) \le \pi^{-1}e^{-\frac14 K_n^2}\Bigr)&, \end{aligned} \tag{15.21}$$

since $\kappa_n \le \frac12 K_n^2$ and $\sup_{h\in C}|h^T\tilde I_{\theta_0,\eta_0}h| \le M^2\|\tilde I_{\theta_0,\eta_0}\| \le \frac14 K_n^2$, for large enough n. Conditioning μ on C, we apply Jensen's inequality to note that, for large enough n,

$$\begin{aligned} P_0^n\Bigl(\int_C e^{h^T\mathbb{G}_n\tilde\ell_{\theta_0,\eta_0}}\,d\mu(h) \le \pi^{-1}e^{-\frac14 K_n^2}\Bigr)& \\ \le P_0^n\Bigl(\int h^T\mathbb{G}_n\tilde\ell_{\theta_0,\eta_0}d\mu(h|C) \le -\tfrac18 K_n^2\Bigr)&, \end{aligned}$$

since $-\log \pi\mu(C) \le \frac18 K_n^2$, for large enough n. The probability on the right is bounded further by Chebyshev's and Jensen's inequalities and can be shown to be of order $O(K_n^{-4})$. Combination with (15.20) and (15.21) then proves (15.19). □

COMPLEMENTS

If h_n is a stochastic sequence, $P^n_{\theta_n(h_n),\eta}f$ denotes the integral,

$$\int f(\omega)\,(dP^n_{\theta_n(h_n(\omega)),\eta}/dP^n_0)(\omega)\,dP^n_0(\omega).$$

Similar considerations apply to Hellinger distances and other integrals involving stochastic (h_n).

Exercises

15.1 In Cox' proportional hazard model (see example 15.2), assume that the survival time T has distribution function F and is absolutely continuous with density $f\colon \mathbb{R} \to [0,\infty)$. Show that the hazard function λ is given by,

$$\lambda(t) = \frac{f(t)}{1-F(t)}.$$

(Hint: apply the Radon-Nikodym theorem.)

15.2 Complete the last steps in the proof of lemma 15.10.

15.3 Argue that marginal consistency can be viewed as consistency in a non-Hausdorff space. More particularly, given a semiparametric model of distributions $P_{\theta,\eta}$, formulate a pseudo-metric to capture marginal consistency.

16

Bayesian efficiency in partial linear regression

Before we prove the semiparametric Bernstein-Von Mises theorem, we consider its application in a regression model. In fact, we apply a simplified version of theorem 15.7 presented in the first section of this lecture.

16.1 A rate-free semiparametric Bernstein-Von Mises theorem

There is room for relaxation of the requirements on model entropy and minimal prior mass, if the limit (15.11) holds in a fixed neighbourhood of η_0. The following corollary applies whenever (15.11) holds for *any rate* (ρ_n). The simplifications are such that the entropy and prior mass conditions become comparable to those for Schwartz' posterior consistency theorem rather than those for posterior rates of convergence.

Corollary 16.1 (Semiparametric Bernstein-Von Mises, rate-free)
Let $X_1, X_2, \ldots$ be distributed i.i.d.-P_0, with $P_0 \in \mathcal{P}$ and let Π_Θ be thick at θ_0. Suppose that for large enough n, the map $h \mapsto s_n(h)$ is continuous P_0^n-almost-surely. Also assume that $\theta \mapsto Q_{\theta,\zeta}$ is stochastically LAN in the θ-direction, for all ζ in an r_H-neighbourhood of $\zeta = 0$ and that the efficient Fisher information $\tilde{I}_{\theta_0 . \eta_0}$ is non-singular. Furthermore, assume that,

(i) For all $\rho > 0$, $N\big(\rho, H, d_H\big) < \infty$ and $\Pi_H\big(K(\rho)\big) > 0$.
(ii) For every $M > 0$, there is an $L > 0$ such that for all $\rho > 0$ and large enough n, $K(\rho) \subset K_n(L\rho, M)$.

and that for every bounded, stochastic (h_n):

(iii) There exists an $r > 0$ such that, $U_n(r, h_n) = O(1)$.
(iv) Hellinger distances satisfy, $\sup_{\eta \in H} H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}) = O(n^{-1/2})$,

and that,

(v) For every (M_n), $M_n \to \infty$, the posterior satisfies,

$$\Pi_n\big(\, \|h\| \le M_n \;\big|\; X_1, \ldots, X_n \,\big) \xrightarrow{P_0} 1.$$

Then the sequence of marginal posteriors for θ converges in total variation to a normal distribution,

$$\sup_A \Big| \Pi_n\big(\, h \in A \;\big|\; X_1, \ldots, X_n \,\big) - N_{\tilde{\Delta}_n, \tilde{I}^{-1}_{\theta_0,\eta_0}}(A) \Big| \xrightarrow{P_0} 0,$$

centred on $\tilde{\Delta}_n$ with covariance matrix $\tilde{I}_{\theta_0,\eta_0}^{-1}$.

Proof Under conditions *(i)*, *(ii)*, *(iv)* and the stochastic LAN assumption, the assertion of corollary 17.3 holds. Due to condition *(iii)*, condition (15.11) is satisfied for large enough n. Condition *(v)* then suffices for the assertion of theorem 17.8. □

A critical note can be made regarding the qualification 'rate-free' of corollary 16.1: although the nuisance rate does not make an explicit appearance, rate restrictions may arise upon further analysis of condition *(v)*. Indeed this is the case in the example of lecture 16, where smoothness requirements on the regression family are interpretable as restrictions on the nuisance rate. However, semiparametric models exist, in which no restrictions on nuisance rates arise in this way: if H is a convex subspace of a linear space and the dependence $\eta \mapsto P_{\theta,\eta}$ is linear (a so-called *convex-linear* model, *e.g.* mixture models, errors-in-variables regression and other information-loss models), the construction of suitable tests, *c.f.* Le Cam (1986) and Birgé (1983, 1984), does not involve Hellinger metric entropy numbers or restrictions on nuisance rates of convergence. Consequently there exists a class of semiparametric examples for which corollary 16.1 stays rate-free even after further analysis of its condition *(v)*.

The particular form of the limiting posterior in theorem 17.8 is a consequence of local asymptotic normality, in this case imposed through (17.16). Other expansions (for instance, in LAN models for non-*i.i.d.* data or under the condition of *local asymptotic exponentiality* (Ibragimov and Has'minskii (1981))) can be dealt with in the same manner if we adapt the limiting form of the posterior accordingly, giving rise to other (*e.g.* one-sided exponential) limit distributions (see Kleijn and Knapik (2012)).

16.2 Partial linear regression

The *partial linear regression* model describes the observation of an *i.i.d.* sample $X_1, X_2, \ldots$ of triplets $X_i = (U_i, V_i, Y_i) \in \mathbb{R}^3$, each assumed to be related through the regression equation,

$$Y = \theta_0 U + \eta_0(V) + e, \tag{16.1}$$

where $e \sim N(0,1)$ is independent of (U,V). Interpreting η_0 as a nuisance parameter, we wish to estimate θ_0. It is assumed that (U,V) has an unknown distribution P, Lebesgue absolutely continuous with density $p\colon \mathbb{R}^2 \to \mathbb{R}$. The distribution P is assumed to be such that $PU = 0$, $PU^2 = 1$ and $PU^4 < \infty$. At a later stage, we also impose $P(U - \mathrm{E}[U|V])^2 > 0$ and a smoothness condition on the conditional expectation $v \mapsto \mathrm{E}[U|V = v]$.

As is well-known Chen (1991), Bickel *et al.* (1998), Mammen and van der Geer (1997), van der Vaart (1998), penalized ML estimation in a smoothness class of regression functions leads to a consistent estimate of the nuisance and efficient point-estimation of the parameter of interest. The necessity of a penalty signals that the choice of a prior for the nuisance is a critical one. Kimeldorf and Wahba (1970) assume that the regression function lies in the Sobolev space $H^k[0,1]$ and define the nuisance prior through the Gaussian process,

$$\eta(t) = \sum_{i=0}^{k} Z_i \frac{t^i}{i!} + (I_{0+}^k W)(t), \tag{16.2}$$

where $W = \{W_t \colon t \in [0,1]\}$ is Brownian motion on $[0,1]$, $(Z_0, \ldots, Z_k)$ form a W-independent, $N(0,1)$-*i.i.d.* sample and I^k_{0+} denotes $(I^1_{0+} f)(t) = \int_0^t f(s)\, ds$, or $I^{i+1}_{0+} f = I^1_{0+}\, I^i_{0+} f$ for all $i \geq 1$. The prior process η is zero-mean Gaussian of (Hölder-)smoothness $k+1/2$ and the resulting posterior mean for η concentrates asymptotically on the smoothing spline that solves the penalized ML problem Wahba (1978), Shen (2002). MCMC simulations based on Gaussian priors have been carried out by Shively, Kohn and Wood (1999).

Here, we reiterate the question how frequentist sufficient conditions are expressed in a Bayesian analysis based on corollary 16.1. We show that with a nuisance of known (Hölder-)smoothness greater than $1/2$, the process (16.2) provides a prior such that the marginal posterior for θ satisfies the Bernstein-Von Mises limit. To facilitate the analysis, we think of the regression function and the process (16.2) as elements of the Banach space $(C[0,1], \|\cdot\|_\infty)$. At a later stage, we relate to Banach subspaces with stronger norms to complete the argument.

Theorem 16.2 *Let $X_1, X_2, \ldots$ be an i.i.d. sample from the partial linear model (16.1) with $P_0 = P_{\theta_0,\eta_0}$ for some $\theta_0 \in \Theta$, $\eta_0 \in H$. Assume that H is a subset of $C[0,1]$ of finite metric entropy with respect to the uniform norm and that H forms a P_0-Donsker class. Regarding the distribution of (U,V), suppose that $PU = 0$, $PU^2 = 1$ and $PU^4 < \infty$, as well as $P(U - \mathrm{E}[U|V])^2 > 0$, $P(U - \mathrm{E}[U|V])^4 < \infty$ and $v \mapsto \mathrm{E}[U|V=v] \in H$. Endow Θ with a prior that is thick at θ_0 and $C[0,1]$ with a prior Π_H such that $H \subset \mathrm{supp}(\Pi_H)$. Then the marginal posterior for θ satisfies the Bernstein-Von Mises limit,*

$$\sup_{B \in \mathscr{B}} \Bigl| \Pi\bigl(\sqrt{n}(\theta - \theta_0) \in B \bigm| X_1, \ldots, X_n \bigr) - N_{\tilde{\Delta}_n, \tilde{I}^{-1}_{\theta_0,f_0}}(B) \Bigr| \xrightarrow{P_0} 0, \tag{16.3}$$

where $\tilde{\ell}_{\theta_0,\eta_0}(X) = e(U - \mathrm{E}[U|V])$ and $\tilde{I}_{\theta_0,\eta_0} = P(U - \mathrm{E}[U|V])^2$.

Proof For any θ and η, $-P_{\theta_0,\eta_0} \log(p_{\theta,\eta}/p_{\theta_0,\eta_0}) = \frac{1}{2} P_{\theta_0,\eta_0}((\theta - \theta_0)U + (\eta - \eta_0)(V))^2$, so that for fixed θ, minimal KL-divergence over H obtains at $\eta^*(\theta) = \eta_0 - (\theta - \theta_0)\, \mathrm{E}[U|V]$, P-almost-surely. For fixed ζ, the submodel $\theta \mapsto Q_{\theta,\zeta}$ satisfies,

$$\begin{aligned} &\log \prod_{i=1}^n \frac{p_{\theta_0 + n^{-1/2}h_n, \eta^*(\theta_0 + n^{-1/2}h_n) + \zeta}}{p_{\theta_0,\eta_0+\zeta}}(X_i) \\ &\quad = \frac{h_n}{\sqrt{n}} \sum_{i=1}^n g_\zeta(X_i) - \tfrac{1}{2} {h_n}^2 P_{\theta_0,\eta_0+\zeta}\, {g_\zeta}^2 + \tfrac{1}{2} {h_n}^2 \bigl(\mathbb{P}_n - P\bigr)(U - \mathrm{E}[U|V])^2, \end{aligned} \tag{16.4}$$

for all stochastic (h_n), with $g_\zeta(X) = e(U - \mathrm{E}[U|V])$, $e = Y - \theta_0 U - (\eta_0 + \zeta)(V) \sim N(0,1)$ under $P_{\theta_0,\eta_0+\zeta}$. Since $PU^2 < \infty$, the last term on the right is $o_{P_{\theta_0,\eta_0+\zeta}}(1)$ if (h_n) is bounded in probability. We conclude that $\theta \mapsto Q_{\theta,\zeta}$ is stochastically LAN. In addition, (16.4) shows that $h \mapsto s_n(h)$ is continuous for every $n \geq 1$. By assumption, $\tilde{I}_{\theta_0,\eta_0} = P_0 {g_0}^2 = P(U - \mathrm{E}[U|V])^2$ is strictly positive. We also observe at this stage that H is totally bounded in $C[0,1]$, so that there exists a constant $D > 0$ such that $\|H\|_\infty \leq D$.

For any $x \in \mathbb{R}^3$ and all ζ, the map $\theta \mapsto \log q_{\theta,\zeta}/q_{\theta_0,\zeta}(x)$ is continuously differentiable on all of Θ, with score $g_{\theta,\zeta}(X) = e(U - \mathrm{E}[U|V]) + (\theta - \theta_0)(U - \mathrm{E}[U|V])^2$. Since $Q_{\theta,\zeta} g^2_{\theta,\zeta} = P(U - \mathrm{E}[U|V])^2 + (\theta - \theta_0)^2 P(U - \mathrm{E}[U|V])^4$ does not depend on ζ and is bounded over $\theta \in [\theta_0 - \rho, \theta_0 + \rho]$, lemma 17.7 says that $U(\rho_n, h_n) = O(1)$ for all $\rho_n \downarrow 0$ and all bounded, stochastic (h_n). So for this model, we can apply the rate-free version of the semiparametric Bernstein-Von Mises theorem, corollary 16.1 and its condition *(iii)* is satisfied.

Regarding condition *(ii)* of corollary 16.1, we first note that, for $M > 0$, $n \geq 1$, $\eta \in H$,

$$\begin{aligned}\sup_{\|h\|\leq M} -\log \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta_0}} = \frac{M^2}{2n}U^2 + \frac{M}{\sqrt{n}}\bigl|U(e-(\eta-\eta_0)(V))\bigr| \\ - e(\eta-\eta_0)(V) + \tfrac{1}{2}(\eta-\eta_0)^2(V),\end{aligned}$$

where $e \sim N(0,1)$ under P_{θ_0,η_0}. With the help of the boundedness of H, the independence of e and (U,V) and the assumptions on the distribution of (U,V), it is then verified that condition *(ii)* of corollary 16.1 holds. Turning to condition *(i)*, it is noted that for all $\eta_1, \eta_2 \in H$, $d_H(\eta_1,\eta_2) \leq -P_{\theta_0,\eta_2}\log(p_{\theta_0,\eta_1}/p_{\theta_0,\eta_2}) = \frac{1}{2}\|\eta_1-\eta_2\|_{2,P}^2 \leq \frac{1}{2}\|\eta_1-\eta_2\|_\infty^2$. Hence, for any $\rho > 0$, $N\bigl(\rho,\mathcal{P}_{\theta_0},d_H\bigr) \leq N\bigl((2\rho)^{1/2},H,\|\cdot\|_\infty\bigr) < \infty$. Similarly, one shows that for all η both $-P_0\log(p_{\theta_0,\eta}/p_{\theta_0,\eta_0})$ and $P_0(\log(p_{\theta_0,\eta}/p_{\theta_0,\eta_0}))^2$ are bounded by $(\frac{1}{2}+D^2)\|\eta-\eta_0\|_\infty^2$. Hence, for any $\rho > 0$, $K(\rho)$ contains a $\|\cdot\|_\infty$-ball. Since $\eta_0 \in \mathrm{supp}(\Pi_H)$, we see that condition *(i)* of corollary 16.1 holds. Noting that $(p_{\theta_n(h),\eta}/p_{\theta_0,\eta}(X))^{1/2} = \exp\left((h/2\sqrt{n})eU - (h^2/4n)U^2\right)$, one derives the η-independent upper bound,

$$H^2\bigl(P_{\theta_n(h_n),\eta},P_{\theta_0,\eta}\bigr) \leq \frac{M^2}{2n}PU^2 + \frac{M^3}{6n^2}PU^4 = O(n^{-1}),$$

for all bounded, stochastic (h_n), so that condition *(iv)* of corollary 16.1 holds.

Concerning condition *(v)*, let (M_n), $M_n \to \infty$ be given and define Θ_n as in section 15.4. Rewrite $\sup_{\eta\in H}\sup_{\theta\in\Theta_n^c}\mathbb{P}_n\log(p_{\theta,\eta}/p_{\theta_0,\eta}) = \sup_{\theta\in\Theta_n^c}((\theta-\theta_0)(\sup_\zeta\mathbb{P}_nZW) - \frac{1}{2}(\theta-\theta_0)^2\,\mathbb{P}_nW^2)$, where $Z = e_0-\zeta(V)$, $W = U-\mathrm{E}[U|V]$. The maximum-likelihood estimate $\hat{\theta}_n$ for θ is therefore of the form $\hat{\theta}_n = \theta_0+R_n$, where $R_n = \sup_\zeta\mathbb{P}_nZW\,/\,\mathbb{P}_nW^2$. Note that $P_0ZW = 0$ and that H is assumed to be P_0-Donsker, so that $\sup_\zeta\mathbb{G}_nZW$ is asymptotically tight. Since in addition, $\mathbb{P}_nW^2 \to P_0W^2$ almost surely and the limit is strictly positive by assumption, $P_0^n(\sqrt{n}\,|R_n| > \frac{1}{4}M_n) = o(1)$. Hence,

$$\begin{aligned}&P_0^n\biggl(\sup_{\eta\in H}\sup_{\theta\in\Theta_n^c}\mathbb{P}_n\log\frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} > -\frac{CM_n^2}{n}\biggr)\\ &\quad\leq P_0^n\biggl(\sup_{\theta\in\Theta_n^c}\Bigl(\tfrac{1}{4}|\theta-\theta_0|\frac{M_n}{n^{1/2}} - \tfrac{1}{2}(\theta-\theta_0)^2\Bigr)\mathbb{P}_nW^2 > -\frac{CM_n^2}{n}\biggr) + o(1)\\ &\quad\leq P_0^n\bigl(\mathbb{P}_nW^2 < 4C\bigr) + o(1).\end{aligned}$$

Since $P_0W^2 > 0$, there exists a $C > 0$ small enough such that the first term on the *r.h.s.* is of order $o(1)$ as well, which shows that condition (15.14) is satisfied. Lemma 15.8 asserts that condition *(v)* of corollary 16.1 is met as well. Assertion 16.3 now holds. □

In the following corollary we choose a prior by picking a suitable k in (16.2) and conditioning on $\|\eta\|_\alpha < M$. The resulting prior is shown to be well-defined below and is denoted $\Pi_{\alpha,M}^k$.

Corollary 16.3 *Let $\alpha > 1/2$ and $M > 0$ be given; choose $H = \{\eta \in C^\alpha[0,1]\colon \|\eta\|_\alpha < M\}$ and assume that $\eta_0 \in C^\alpha[0,1]$. Suppose the distribution of the covariates (U,V) is* *as in theorem 16.2. Then, for any integer $k > \alpha - 1/2$, the conditioned prior $\Pi^k_{\alpha,M}$ is well-defined and gives rise to a marginal posterior for θ satisfying (16.3).*

Proof Choose k as indicated; the Gaussian distribution of η over $C[0,1]$ is based on the RKHS $H^{k+1}[0,1]$ and denoted Π^k. Since η in (16.2) has smoothness $k+1/2 > \alpha$, $\Pi^k(\eta \in C^\alpha[0,1]) = 1$. Hence, one may also view η as a Gaussian element in the Hölder class $C^\alpha[0,1]$, which forms a separable Banach space even with strengthened norm $\|\cdot\| = \|\eta\|_\infty + \|\cdot\|_\alpha$, without changing the RKHS. The trivial embedding of $C^\alpha[0,1]$ into $C[0,1]$ is one-to-one and continuous, enabling identification of the prior induced by η on $C^\alpha[0,1]$ with the prior Π^k on $C[0,1]$. Given $\eta_0 \in C^\alpha[0,1]$ and a sufficiently smooth kernel ϕ_σ with bandwidth $\sigma > 0$, consider $\phi_\sigma \star \eta_0 \in H^{k+1}[0,1]$. Since $\|\eta_0 - \phi_\sigma \star \eta_0\|_\infty$ is of order σ^α and a similar bound exists for the α-norm of the difference, η_0 lies in the closure of the RKHS both with respect to $\|\cdot\|_\infty$ and to $\|\cdot\|$. Particularly, η_0 lies in the support of Π^k, in $C^\alpha[0,1]$ with norm $\|\cdot\|$. Hence, $\|\cdot\|$-balls centred on η_0 receive non-zero prior mass, *i.e.* $\Pi^k(\|\eta - \eta_0\| < \rho) > 0$ for all $\rho > 0$. Therefore, $\Pi^k(\|\eta - \eta_0\|_\infty < \rho, \|\eta\|_\alpha < \|\eta_0\|_\alpha + \rho) > 0$, which guarantees that $\Pi^k(\|\eta - \eta_0\|_\infty < \rho, \|\eta\|_\alpha < M) > 0$, for small enough $\rho > 0$. This implies that $\Pi^k(\|\eta\|_\alpha < M) > 0$ and,

$$\Pi^k_{\alpha,M}(B) = \Pi^k\big(\, B \mid \|\eta\|_\alpha < M\,\big),$$

is well-defined for all Borel-measurable $B \subset C[0,1]$. Moreover, it follows that $\Pi^k_{\alpha,M}(\|\eta - \eta_0\|_\infty < \rho) > 0$ for all $\rho > 0$. We conclude that k times integrated Brownian motion started at random, conditioned to be bounded by M in α-norm, gives rise to a prior that satisfies $\mathrm{supp}(\Pi^k_{\alpha,M}) = H$. As is well-known van der Vaart and Wellner (1996), the entropy numbers of H with respect to the uniform norm satisfy, for every $\rho > 0$, $N(\rho, H, \|\cdot\|_\infty) \le K\rho^{-1/\alpha}$, for some constant $K > 0$ that depends only on α and M. The associated bound on the bracketing entropy gives rise to finite bracketing integrals, so that H universally Donsker. Then, if the distribution of the covariates (U, V) is as assumed in theorem 16.2, the Bernstein-Von Mises limit (16.3) holds. □

Exercises

16.1 Show that the efficient score function for the partial linear model is given by $\tilde{\ell}_{\theta_0,\eta_0}(X) = e(U - \mathrm{E}[U|V])$.

16.2 Speculate concerning the question whether the condition $\|\eta\|_\alpha \le M$ is a necessary condition.

17

The proof of the semiparametric Bernstein-von Mises theorem

In this lecture, we consider the proof of the semiparametric Bernstein-Von Mises theorem, broken up in three major parts. The first step concerns a proof of consistency for the nuisance parameter under perturbation of the parameter of interest of size proportional to $n^{-1/2}$. Building on that, the second step shows under which conditions integrals of likelihoods with respect to the nuisance prior display the LAN property. The last step uses that LAN expansion of integrated likelihoods to demonstrate asymptotic normality of the marginal posterior distribution for the parameter of interest, with a proof that closely resembles that of theorem 14.3.

17.1 Posterior convergence under perturbation

We consider contraction of the posterior around least-favourable submodels. We express this form of posterior convergence by showing that (under suitable conditions) the conditional posterior for the nuisance parameter contracts around the least-favourable submodel, conditioned on a sequence $\theta_n(h_n)$ for the parameter of interest with $h_n = O_{P_o}(1)$. We view the sequence of models $\mathcal{P}_{\theta_n(h_n)}$ as a random perturbation of the model $\mathcal{P}_{\theta_0}$ and generalize Ghosal *et al.* (2000) to describe posterior contraction. Ultimately, random perturbation of θ represents the 'appropriate form of uniformity' referred to just after definition (15.8). Given a rate sequence (ρ_n), $\rho_n \downarrow 0$, we say that the conditioned nuisance posterior is *consistent under $n^{-1/2}$-perturbation at rate ρ_n*, if,

$$\Pi_n\big(\, D^c(\theta,\rho_n) \;\big|\; \theta = \theta_0 + n^{-1/2}h_n\, ;\, X_1,\ldots,X_n\,\big) \xrightarrow{P_0} 0, \qquad (17.1)$$

for all bounded, stochastic sequences (h_n).

Theorem 17.1 (Posterior rate of convergence under perturbation)
Assume that there exists a sequence (ρ_n) with $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$ such that for all $M > 0$ and every bounded, stochastic (h_n):

(i) There exists a constant $K > 0$ such that for large enough n,

$$\Pi_H\big(K_n(\rho_n, M)\big) \geq e^{-Kn\rho_n^2}. \qquad (17.2)$$

(ii) For $L > 0$ large enough, there exist (ϕ_n) such that for large enough n,

$$P_0^n\phi_n \to 0, \qquad \sup_{\eta\in D^c(\theta_0,L\rho_n)} P^n_{\theta_n(h_n),\eta}(1-\phi_n) \leq e^{-\frac{1}{4}L^2n\rho_n^2}. \qquad (17.3)$$

(iii) The least-favourable submodel satisfies $d_H(\eta^*(\theta_n(h_n)),\eta_0) = o(\rho_n)$.

Then, for every bounded, stochastic (h_n) *there exists an* $L > 0$ *such that the conditional nuisance posterior converges as,*

$$\Pi\big(D^c(\theta, L\rho_n) \bigm| \theta = \theta_0 + n^{-1/2}h_n;\, X_1, \ldots, X_n \big) = o_{P_0}(1), \tag{17.4}$$

under $n^{-1/2}$*-perturbation.*

Proof Let (h_n) be a stochastic sequence bounded by M and let $0 < C < 1$ be given. Let K and (ρ_n) be as in conditions *(i)* and *(ii)*. Choose $L > 4\sqrt{1+K+C}$ and large enough to satisfy condition *(ii)* for some (ϕ_n). By lemma 17.4, the events,

$$A_n = \left\{ \int_H \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta) \geq e^{-(1+C)n\rho_n^2}\, \Pi_H(K_n(\rho_n, M)) \right\},$$

satisfy $P_0^n(A_n^c) \to 0$. Using also the first limit in (17.3), we then derive,

$$\begin{aligned} &P_0^n \Pi\big(D^c(\theta, L\rho_n) \bigm| \theta = \theta_n(h_n);\, X_1, \ldots, X_n \big) \\ &\qquad \leq P_0^n \Pi\big(D^c(\theta, L\rho_n) \bigm| \theta = \theta_n(h_n);\, X_1, \ldots, X_n \big)\, 1_{A_n}\, (1-\phi_n) + o(1), \end{aligned}$$

(even with random (h_n), the posterior $\Pi(\,\cdot\,|\theta = \theta_n(h_n);\, X_1, \ldots, X_n\big) \leq 1$, by definition (15.5)). The first term on the *r.h.s.* can be bounded further by the definition of the events A_n,

$$\begin{aligned} &P_0^n \Pi\big(D^c(\theta, L\rho_n) \bigm| \theta = \theta_n;\, X_1, \ldots, X_n \big)\, 1_{A_n}\, (1-\phi_n) \\ &\qquad \leq \frac{e^{(1+C)n\rho_n^2}}{\Pi_H(K_n(\rho_n, M))} P_0^n \left(\int_{D^c(\theta_n(h_n), L\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)\, (1-\phi_n)\, d\Pi_H \right). \end{aligned}$$

Due to condition *(iii)* it follows that,

$$D(\theta_0, \tfrac{1}{2}L\rho_n) \subset \bigcap_{n \geq 1} D(\theta_n(h_n), L\rho_n), \tag{17.5}$$

for large enough n. Therefore,

$$\begin{aligned} &P_0^n \int_{D^c(\theta_n(h_n), L\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)\, (1-\phi_n)\, d\Pi_H(\eta) \\ &\qquad\qquad \leq \int_{D^c(\theta_0, \frac{1}{2}L\rho_n)} P^n_{\theta_n(h_n),\eta}(1-\phi_n)\, d\Pi_H(\eta). \end{aligned} \tag{17.6}$$

Upon substitution of (17.6) and with the use of the second bound in (17.3) and (17.2), the choice we made earlier for L proves the assertion. □

We conclude from the above that besides sufficiency of prior mass, the crucial condition for consistency under perturbation is the existence of a test sequence (ϕ_n) satisfying (17.3). To find sufficient conditions, we follow a construction of tests based on the Hellinger geometry of the model, generalizing the approach of Birgé (1983, 1984) and Le Cam (1986) to $n^{-1/2}$-perturbed context. It is easiest to illustrate their approach by considering the problem of testing/estimating η when θ_0 is known: we cover the nuisance model $\{P_{\theta_0,\eta}\colon \eta \in H\}$ by a minimal collection of Hellinger balls B of radii (ρ_n), each of which is convex and hence testable against P_0 with power bounded by $\exp(-\frac{1}{4}\, n\, H^2(P_0, B))$, based on the minimax theorem. The tests for the covering Hellinger balls are combined into a single test for the non-convex alternative $\{P\colon H(P,P_0) \geq \rho_n\}$ against P_0. The order of the cover controls the power of the combined test. Therefore the construction requires an upper bound to Hellinger metric entropy numbers,

$$N\big(\rho_n, \mathcal{P}_{\theta_0}, H\big) \leq e^{n\rho_n^2}, \tag{17.7}$$

which is interpreted as indicative of the nuisance model's complexity in the sense that the lower bound to the collection of rates (ρ_n) solving (17.7), is the Hellinger minimax rate for estimation of η_0. In the $n^{-1/2}$-perturbed problem, the alternative does not just consist of the complement of a Hellinger-ball in the nuisance factor H, but also has an extent in the θ-direction shrinking at rate $n^{-1/2}$. Condition (17.8) below guarantees that Hellinger covers of H like the above are large enough to accommodate the θ-extent of the alternative, the implication being that the test sequence one constructs for the nuisance in case θ_0 is known, can also be used when θ_0 is known only up to $n^{-1/2}$-perturbation. Therefore, the entropy bound in lemma 17.2 is (17.7). Geometrically, (17.8) requires that $n^{-1/2}$-perturbed versions of the nuisance model are contained in a narrowing sequence of metric cones based at P_0. In differentiable models, the Hellinger distance $H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})$ is typically of order $O(n^{-1/2})$ for all $\eta \in H$. So if, in addition, $n\rho_n^2 \to \infty$, limit (17.8) is expected to hold pointwise in η. Then only the uniform character of (17.8) truly forms a condition.

Lemma 17.2 (Testing under perturbation)
If (ρ_n) satisfies $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$ and the following requirements are met:

(i) For all n large enough, $N\big(\rho_n, H, d_H\big) \leq e^{n\rho_n^2}$.

(ii) For all $L > 0$ and all bounded, stochastic (h_n),

$$\sup_{\{\eta \in H: d_H(\eta,\eta_0) \geq L\rho_n\}} \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} = o(1). \tag{17.8}$$

Then for all $L \geq 4$, there exists a test sequence (ϕ_n) such that for all bounded, stochastic (h_n),

$$P_0^n \phi_n \to 0, \qquad \sup_{\eta \in D^c(\theta_0, L\rho_n)} P^n_{\theta_n(h_n),\eta}(1-\phi_n) \leq e^{-\frac{1}{4}L^2 n\rho_n^2}, \tag{17.9}$$

for large enough n.

Proof Let (ρ_n) be such that *(i)–(ii)* are satisfied. Let (h_n) and $L \geq 4$ be given. For all $j \geq 1$, define $H_{j,n} = \{\eta \in H\colon jL\rho_n \leq d_H(\eta_0,\eta) \leq (j+1)L\rho_n\}$ and $\mathcal{P}_{j,n} = \{P_{\theta_0,\eta}\colon \eta \in H_{j,n}\}$. Cover $\mathcal{P}_{j,n}$ with Hellinger balls $B_{i,j,n}(\frac{1}{4}jL\rho_n)$, where,

$$B_{i,j,n}(r) = \big\{P\colon H(P_{i,j,n}, P) \leq r\big\},$$

and $P_{i.j.n} \in \mathcal{P}_{j,n}$, *i.e.* there exists an $\eta_{i,j,n} \in H_{j,n}$ such that $P_{i,j,n} = P_{\theta_0,\eta_{i,j,n}}$. Denote $H_{i,j,n} = \{\eta \in H_{j,n}\colon P_{\theta_0,\eta} \in B_{i,j,n}(\frac{1}{4}jL\rho_n)\}$. By assumption, the minimal number of such balls needed to cover $\mathcal{P}_{i,j}$ is finite; we denote the corresponding covering number by $N_{j,n}$, *i.e.* $1 \leq i \leq N_{j,n}$.

Let $\eta \in H_{j,n}$ be given. There exists an i ($1 \leq i \leq N_{j,n}$) such that $d_H(\eta, \eta_{i,j,n}) \leq$

$\frac{1}{4}jL\rho_n$. Then, by the triangle inequality, the definition of $H_{j,n}$ and assumption (17.8),

$$\begin{aligned} H\big(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta_{i,j,n}}\big) &\leq H\big(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}\big) + H\big(P_{\theta_0,\eta}, P_{\theta_0,\eta_{i,j,n}}\big) \\ &\leq \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} H\big(P_{\theta_0,\eta}, P_0\big) + \tfrac{1}{4}jL\rho_n \\ &\leq \left(\sup_{\{\eta\in H: d_H(\eta,\eta_0)\geq L\rho_n\}} \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} \right)(j+1)L\rho_n + \tfrac{1}{4}jL\rho_n \\ &\leq \tfrac{1}{2}jL\rho_n, \end{aligned} \tag{17.10}$$

for large enough n. We conclude that there exists an $N \geq 1$ such that for all $n \geq N$, $j \geq 1$, $1 \leq i \leq N_{j,n}$, $\eta \in H_{i,j,n}$, $P_{\theta_n(h_n),\eta} \in B_{i,j,n}(\frac{1}{2}jL\rho_n)$. Moreover, Hellinger balls are convex and for all $P \in B_{i,j,n}(\frac{1}{2}jL\rho_n)$, $H(P,P_0) \geq \frac{1}{2}jL\rho_n$. As a consequence of the minimax theorem, there exists a test sequence $(\phi_{i,j,n})_{n\geq 1}$ such that,

$$P_0^n\phi_{i,j,n} \vee \sup_P P^n(1-\phi_{i,j,n}) \leq e^{-nH^2(B_{i,j,n}(\frac{1}{2}jL\rho_n),P_0)} \leq e^{-\frac{1}{4}nj^2L^2\rho_n^2},$$

where the supremum runs over all $P \in B_{i,j,n}(\frac{1}{2}jL\rho_n)$. Defining, for all $n \geq 1$, $\phi_n = \sup_{j\geq 1}\max_{1\leq i\leq N_{j,n}} \phi_{i,j,n}$, we find (for details, see the proof of theorem 3.10 in Kleijn (2003)) that,

$$P_0^n\phi_n \leq \sum_{j\geq 1} N_{j,n} e^{-\frac{1}{4}L^2j^2n\rho_n^2}, \qquad P^n(1-\phi_n) \leq e^{-\frac{1}{4}L^2n\rho_n^2}, \tag{17.11}$$

for all $P = P_{\theta_n(h_n),\eta}$ and $\eta \in D^c(\theta_0, L\rho_n)$. Since $L \geq 4$, we have for all $j \geq 1$,

$$N_{j,n} = N\big(\tfrac{1}{4}Lj\rho_n, \mathcal{P}_{j,n}, H\big) \leq N\big(\tfrac{1}{4}Lj\rho_n, \mathcal{P}, H\big) \leq N(\rho_n, \mathcal{P}, H) \leq e^{n\rho_n^2}, \tag{17.12}$$

by assumption (17.7). Upon substitution of (17.12) into (17.11), we obtain the following bounds,

$$P_0^n\phi_n \leq \frac{e^{(1-\frac{1}{4}L^2)n\rho_n^2}}{1-e^{-\frac{1}{4}L^2n\rho_n^2}}, \qquad \sup_{\eta\in D^c(\theta_0,L\rho_n)} P^n_{\theta_n(h_n),\eta}(1-\phi_n) \leq e^{-\frac{1}{4}L^2n\rho_n^2},$$

for large enough n, which implies assertion (17.9). □

In preparation of corollary 16.1, we also provide a version of theorem 17.1 that only asserts consistency under $n^{-1/2}$-perturbation at *some* rate while relaxing bounds for prior mass and entropy. In the statement of the corollary, we make use of the family of Kullback-Leibler neighbourhoods that would play a role for the posterior of the nuisance if θ_0 were known:

$$K(\rho) = \Big\{\eta \in H : -P_0 \log\frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}} \leq \rho^2, P_0\Big(\log\frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}\Big)^2 \leq \rho^2\Big\}, \tag{17.13}$$

for all $\rho > 0$. The proof below follows steps similar to those in the proof of corollary 2.1 in Kleijn and van der Vaart (2006).

Corollary 17.3 (Posterior consistency under perturbation)
Assume that for all $\rho > 0$, $N\big(\rho, H, d_H\big) < \infty$, $\Pi_H(K(\rho)) > 0$ *and,*

(i) For all $M > 0$ there is an $L > 0$ such that for all $\rho > 0$ and large enough n, $K(\rho) \subset K_n(L\rho, M)$.

(ii) For every bounded random sequence (h_n), the quantity $\sup_{\eta \in H} H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})$ and $H(P_{\theta_0,\eta^*(\theta_n(h_n))}, P_{\theta_0,\eta_0})$ are of order $O(n^{-1/2})$.

Then there exists a sequence (ρ_n), $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$, such that the conditional nuisance posterior converges under $n^{-1/2}$-perturbation at rate (ρ_n).

Proof We follow the proof of corollary 2.1 in Kleijn and van der Vaart (2006) and add that, under condition *(ii)*, (17.8) and condition *(iii)* of theorem 17.1 are satisfied. We conclude that there exists a test sequence satisfying (17.3). Then, the assertion of theorem 17.1 holds. □

The following lemma generalizes lemma 8.1 in Ghosal *et al.* (2000) to the $n^{-1/2}$-perturbed setting.

Lemma 17.4 *Let (h_n) be stochastic and bounded by some $M > 0$. Then,*

$$P_0^n\biggl(\int_H \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta) < e^{-(1+C)n\rho^2}\, \Pi_H(K_n(\rho, M))\biggr) \leq \frac{1}{C^2 n\rho^2}, \tag{17.14}$$

for all $C > 0$, $\rho > 0$ and $n \geq 1$.

Proof See the proof of lemma 8.1 in Ghosal *et al.* (2000) (dominating the h_n-dependent log-likelihood ratio immediately after the first application of Jensen's inequality). □

17.2 Integrating local asymptotic normality

The smoothness condition in the Le Cam's parametric Bernstein-Von Mises theorem is a LAN expansion of the likelihood, which is replaced in semiparametric context by a stochastic LAN expansion of the integrated likelihood (15.9). In this section, we consider sufficient conditions under which the localized integrated likelihood,

$$s_n(h) = \int_H \prod_{i=1}^n \frac{p_{\theta_0+n^{-1/2}h,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta), \tag{17.15}$$

has the *integral LAN* property, *i.e.* s_n allows an expansion of the form,

$$\log \frac{s_n(h_n)}{s_n(0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} h_n^T \tilde{\ell}_{\theta_0,\eta_0} - \tfrac{1}{2} h_n^T \tilde{I}_{\theta_0,\eta_0} h_n + o_{P_0}(1), \tag{17.16}$$

for every random sequence $(h_n) \subset \mathbb{R}^k$ of order $O_{P_0}(1)$, as required in theorem 17.8. Theorem 17.6 assumes that the model is stochastically LAN and requires consistency under $n^{-1/2}$-perturbation for the nuisance posterior. Consistency not only allows us to restrict sufficient conditions to neighbourhoods of η_0 in H, but also enables lifting of the LAN expansion of the integrand in (17.15) to an expansion of the integral s_n itself, *c.f.* (17.16). The posterior concentrates on the least-favourable submodel so that only the least-favourable expansion at η_0 contributes to (17.16) asymptotically. For this reason, the intergral LAN expansion is determined by the efficient score function (and not some other influence function). Ultimately, occurrence of the efficient score lends the marginal posterior (and statistics based upon it) properties of frequentist semiparametric optimality.

To derive theorem 17.6, we reparametrize the model *c.f.* (15.10). While yielding adaptivity, this reparametrization also leads to θ-dependence in the prior for ζ, a technical issue that we tackle before addressing the main point of this section. We show that the prior mass of the relevant neighbourhoods displays the appropriate type of stability, under a condition on local behaviour of Hellinger distances in the least-favourable model. For smooth least-favourable submodels, typically $d_H(\eta^*(\theta_n(h_n)),\eta_0) = O(n^{-1/2})$ for all bounded, stochastic (h_n), which suffices.

Lemma 17.5 (Prior stability)
Let (h_n) be a bounded, stochastic sequence of perturbations and let Π_H be any prior on H. Let (ρ_n) be such that $d_H\bigl(\eta^*(\theta_n(h_n)),\eta_0\bigr) = o(\rho_n)$. Then the prior mass of radius-ρ_n neighbourhoods of η^* is stable, i.e.,

$$\Pi_H\bigl(D(\theta_n(h_n),\rho_n)\bigr) = \Pi_H\bigl(D(\theta_0,\rho_n)\bigr) + o(1). \tag{17.17}$$

Proof Let (h_n) and (ρ_n) be such that $d_H\bigl(\eta^*(\theta_n(h_n)),\eta_0\bigr) = o(\rho_n)$. Denote $D(\theta_n(h_n),\rho_n)$ by D_n and $D(\theta_0,\rho_n)$ by C_n for all $n\geq 1$. Since,

$$\Bigl|\Pi_H(D_n) - \Pi_H(C_n)\Bigr| \leq \Pi_H\bigl((D_n\cup C_n) - (D_n\cap C_n)\bigr),$$

we consider the sequence of symmetric differences. Fix some $0<\alpha<1$. Then for all $\eta\in D_n$ and all n large enough, $d_H(\eta,\eta_0)\leq d_H(\eta,\eta^*(\theta_n(h_n))) + d_H(\eta^*(\theta_n(h_n)),\eta_0)\leq (1+\alpha)\rho_n$, so that $D_n\cup C_n\subset D(\theta_0,(1+\alpha)\rho_n)$. Furthermore, for large enough n and any $\eta\in D(\theta_0,(1-\alpha)\rho_n)$, $d_H(\eta,\eta^*(\theta_n(h_n)))\leq d_H(\eta,\eta_0)+d_H(\eta_0,\eta^*(\theta_n(h_n)))\leq \rho_n + d_H(\eta_0,\eta^*(\theta_n(h_n))) - \alpha\rho_n < \rho_n$, so that $D(\theta_0,(1-\alpha)\rho_n)\subset D_n\cap C_n$. Therefore,

$$(D_n\cup C_n) - (D_n\cap C_n) \subset D(\theta_0,(1+\alpha)\rho_n)\bigr) - D(\theta_0,(1-\alpha)\rho_n) \to \emptyset,$$

which implies (17.17). □

Once stability of the nuisance prior is established, theorem 17.6 hinges on stochastic local asymptotic normality of the submodels $t\mapsto Q_{\theta_0+t,\zeta}$, for all ζ in an r_H-neighbourhood of $\zeta=0$. We assume there exists a $g_\zeta\in L_2(Q_{\theta_0,\zeta})$ such that for every random (h_n) bounded in $Q_{\theta_0,\zeta}$-probability,

$$\log\prod_{i=1}^{n}\frac{q_{\theta+n^{-1/2}h_n,\zeta}}{q_{\theta_0,0}}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_n^T g_\zeta(X_i) - \tfrac{1}{2}\,h_n^T I_\zeta h_n + R_n(h_n,\zeta), \tag{17.18}$$

where $I_\zeta = Q_{\theta_0,\zeta}g_\zeta g_\zeta^T$ and $R_n(h_n,\zeta) = o_{Q_{\theta_0,\zeta}}(1)$. Equation (17.18) specifies the tangent set with respect to which differentiability of the model is required. Note that $g_0 = \tilde{\ell}_{\theta_0,\eta_0}$.

Theorem 17.6 (Integral local asymptotic normality)
Suppose that $\theta\mapsto Q_{\theta,\zeta}$ is stochastically LAN for all ζ in an r_H-neighbourhood of $\zeta=0$. Furthermore, assume that posterior consistency under $n^{-1/2}$-perturbation obtains with a rate (ρ_n) also valid in (15.11). Then the integral LAN-expansion (17.16) holds.

Proof Throughout this proof $G_n(h,\zeta) = \sqrt{n}\,h^T\mathbb{P}_n g_\zeta - \frac{1}{2}h^T I_\zeta h$, for all h and all ζ. Furthermore, we abbreviate $\theta_n(h_n)$ to θ_n and omit explicit notation for $(X_1,\ldots,X_n)$-dependence in several places.

Let $\delta,\epsilon > 0$ be given and let $\theta_n = \theta_0 + n^{-1/2}h_n$ with (h_n) bounded in P_0-probability. Then there exists a constant $M > 0$ such that $P_0^n(\|h_n\| > M) < \frac{1}{2}\delta$ for all $n \geq 1$. With (h_n) bounded, the assumption of consistency under $n^{-1/2}$-perturbation says that,

$$P_0^n\Big(\log\Pi\big(\,D(\theta,\rho_n)\;\big|\;\theta=\theta_n\,;\,X_1,\ldots,X_n\,\big)\geq -\epsilon\Big) > 1-\tfrac{1}{2}\delta.$$

for large enough n. This implies that the posterior's numerator and denominator are related through,

$$\begin{aligned}P_0^n\bigg(\int_H\prod_{i=1}^n\frac{p_{\theta_n,\eta}}{p_{\theta_0,\eta_0}}(X_i)\,d\Pi_H(\eta)&\\ \leq e^\epsilon\,1_{\{\|h_n\|\leq M\}}\int_{D(\theta_n,\rho_n)}\prod_{i=1}^n&\frac{p_{\theta_n,\eta}}{p_{\theta_0,\eta_0}}(X_i)\,d\Pi_H(\eta)\bigg) > 1-\delta.\end{aligned}\tag{17.19}$$

We continue with the integral over $D(\theta_n,\rho_n)$ under the restriction $\|h_n\|\leq M$ and parametrize the model locally in terms of (θ,ζ) (see (15.10)):

$$\int_{D(\theta_n,\rho_n)}\prod_{i=1}^n\frac{p_{\theta_n,\eta}}{p_{\theta_0,\eta_0}}(X_i)\,d\Pi_H(\eta) = \int_{B(\rho_n)}\prod_{i=1}^n\frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi\big(\,\zeta\;\big|\;\theta=\theta_n\big),\tag{17.20}$$

where $\Pi(\,\cdot\,|\,\theta\,)$ denotes the prior for ζ given θ, *i.e.* Π_H translated over $\eta^*(\theta)$. Next we note that by Fubini's theorem and the domination condition (15.11), there exists a constant $L > 0$ such that,

$$\begin{aligned}\bigg|\,P_0^n\int_{B(\rho_n)}\prod_{i=1}^n\frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,\big(d\Pi(\zeta\mid\theta=\theta_n)-d\Pi(\zeta\mid\theta=\theta_0)\big)\bigg|&\\ \leq L\,\Big|\,\Pi\big(\,B(\rho_n)\;\big|\;\theta=\theta_n\,\big)-\Pi\big(\,B(\rho_n)\;\big|\;\theta=\theta_0\,\big)\Big|&,\end{aligned}$$

for large enough n. Since the least-favourable submodel is stochastically LAN, lemma 17.5 asserts that the difference on the *r.h.s.* of the above display is $o(1)$, so that,

$$\int_{B(\rho_n)}\prod_{i=1}^n\frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi\big(\,\zeta\;\big|\;\theta=\theta_n\big) = \int_{B(\rho_n)}\prod_{i=1}^n\frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi(\zeta)+o_{P_0}(1),\tag{17.21}$$

where we use the notation $\Pi(A) = \Pi(\,\zeta\in A\,|\,\theta=\theta_0\,)$ for brevity. We define for all ζ, $\epsilon > 0$, $n \geq 1$ the events $F_n(\zeta,\epsilon) = \big\{\sup_h|G_n(h,\zeta)-G_n(h,0)|\leq\epsilon\big\}$. With (15.11) as a domination condition, Fatou's lemma and the fact that $F_n^c(0,\epsilon)=\emptyset$ lead to,

$$\begin{aligned}\limsup_{n\to\infty}\int_{B(\rho_n)}&Q^n_{\theta_n,\zeta}\big(F_n^c(\zeta,\epsilon)\big)\,d\Pi(\zeta)\\ &\leq\int\limsup_{n\to\infty}1_{B(\rho_n)-\{0\}}(\zeta)\,Q^n_{\theta_n,\zeta}\big(F_n^c(\zeta,\epsilon)\big)\,d\Pi(\zeta)=0,\end{aligned}\tag{17.22}$$

(again using (15.11) in the last step). Combined with Fubini's theorem, this suffices to conclude that,

$$\int_{B(\rho_n)} \prod_{i=1}^{n} \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi(\zeta) = \int_{B(\rho_n)} \prod_{i=1}^{n} \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) 1_{F_n(\zeta,\epsilon)}\,d\Pi(\zeta) + o_{P_0}(1), \qquad (17.23)$$

and we continue with the first term on the *r.h.s.*. By stochastic local asymptotic normality for every ζ, expansion (17.18) of the log-likelihood implies that,

$$\prod_{i=1}^{n} \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) = \prod_{i=1}^{n} \frac{q_{\theta_0,\zeta}}{q_{\theta_0,0}}(X_i)\, e^{G_n(h_n,\zeta)+R_n(h_n,\zeta)}, \qquad (17.24)$$

where the rest term is of order $o_{Q_{\theta_0,\zeta}}(1)$. Accordingly, we define, for every ζ, the events $A_n(\zeta,\epsilon) = \{|R_n(h_n,\zeta)| \le \frac{1}{2}\epsilon\}$, so that $Q^n_{\theta_0,\zeta}(A^c_n(\zeta,\epsilon)) \to 0$. Contiguity then implies that $Q^n_{\theta_n,\zeta}(A^c_n(\zeta,\epsilon)) \to 0$ as well. Reasoning as in (17.23) we see that,

$$\begin{aligned} &\int_{B(\rho_n)} \prod_{i=1}^{n} \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\, 1_{F_n(\zeta,\epsilon)}\,d\Pi(\zeta) \\ &\qquad = \int_{B(\rho_n)} \prod_{i=1}^{n} \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\, 1_{A_n(\zeta,\epsilon)\cap F_n(\zeta,\epsilon)}\,d\Pi(\zeta) + o_{P_0}(1). \end{aligned} \qquad (17.25)$$

For fixed n and ζ and for all $(X_1,\ldots,X_n) \in A_n(\zeta,\epsilon)\cap F_n(\zeta,\epsilon)$:

$$\left| \log \prod_{i=1}^{n} \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) - G_n(h_n,0) \right| \le 2\epsilon,$$

so that the first term on the *r.h.s.* of (17.25) satisfies the bounds,

$$\begin{aligned} e^{G_n(h_n,0)-2\epsilon} &\int_{B(\rho_n)} \prod_{i=1}^{n} \frac{q_{\theta_0,\zeta}}{q_{\theta_0,0}}(X_i)\, 1_{A_n(\zeta,\epsilon)\cap F_n(\zeta,\epsilon)}\,d\Pi(\zeta) \\ &\le \int_{B(\rho_n)} \prod_{i=1}^{n} \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\, 1_{A_n(\zeta,\epsilon)\cap F_n(\zeta,\epsilon)}\,d\Pi(\zeta) \\ &\le e^{G_n(h_n,0)+2\epsilon} \int_{B(\rho_n)} \prod_{i=1}^{n} \frac{q_{\theta_0,\zeta}}{q_{\theta_0,0}}(X_i)\, 1_{A_n(\zeta,\epsilon)\cap F_n(\zeta,\epsilon)}\,d\Pi(\zeta). \end{aligned} \qquad (17.26)$$

The integral factored into lower and upper bounds can be relieved of the indicator for $A_n \cap F_n$ by reversing the argument that led to (17.23) and (17.25) (with θ_0 replacing θ_n), at the expense of an $e^{o_{P_0}(1)}$-factor. Substituting in (17.26) and using, consecutively, (17.25), (17.23), (17.21) and (17.19) for the bounded integral, we find,

$$e^{G_n(h_n,0)-3\epsilon+o_{P_0}(1)}\, s_n(0) \le s_n(h_n) \le e^{G_n(h_n,0)+3\epsilon+o_{P_0}(1)} s_n(0).$$

Since this holds with arbitrarily small $0 < \epsilon' < \epsilon$ for large enough n, it proves (17.16). □

With regard to the nuisance rate (ρ_n), we first note that our proof of theorem 15.7 fails if the slowest rate required to satisfy (15.11) vanishes *faster* then the optimal rate for convergence under $n^{-1/2}$-perturbation (as determined in (17.7) and (17.2)).

However, the rate (ρ_n) does not appear in assertion (17.16), so if said contradiction between conditions (15.11) and (17.7)/(17.2) does not occur, the sequence (ρ_n) can remain entirely internal to the proof of theorem 17.6. More particularly, if condition (15.11) holds for *any* (ρ_n) such that $n\rho_n^2 \to \infty$, integral LAN only requires consistency under $n^{-1/2}$-perturbation at *some* such (ρ_n). In that case, we may appeal to corollary 17.3 instead of theorem 17.1, thus relaxing conditions on model entropy and nuisance prior. The following lemma shows that a first-order Taylor expansion of likelihood ratios combined with a boundedness condition on certain Fisher information coefficients is enough to enable use of corollary 17.3 instead of theorem 17.1.

Lemma 17.7 *Let Θ be one-dimensional. Assume that there exists a $\rho > 0$ such that for every $\zeta \in B(\rho)$ and all x in the samplespace, the map $\theta \mapsto \log(q_{\theta,\zeta}/q_{\theta_0,\zeta})(x)$ is continuously differentiable on $[\theta_0 - \rho, \theta_0 + \rho]$ with Lebesgue-integrable derivative $g_{\theta,\zeta}(x)$ such that,*

$$\sup_{\zeta \in B(\rho)} \sup_{\{\theta : |\theta - \theta_0| < \rho\}} Q_{\theta,\zeta} g_{\theta,\zeta}^2 < \infty. \tag{17.27}$$

Then, for every $\rho_n \downarrow 0$ and all bounded, stochastic (h_n), $U_n(\rho_n, h_n) = O(1)$.

Proof Let (h_n) be stochastic and upper-bounded by $M > 0$. For every ζ and all $n \geq 1$,

$$\begin{aligned} Q^n_{\theta_0,\zeta}\biggl|\prod_{i=1}^n \frac{q_{\theta_n(h_n),\zeta}}{q_{\theta_0,\zeta}}(X_i) - 1\biggr| &= Q^n_{\theta_0,\zeta}\biggl|\int_{\theta_0}^{\theta_n(h_n)} \sum_{i=1}^n g_{\theta',\zeta}(X_i) \prod_{j=1}^n \frac{q_{\theta',\zeta}}{q_{\theta_0,\zeta}}(X_j)\, d\theta'\biggr| \\ &\leq \int_{\theta_0 - \frac{M}{\sqrt{n}}}^{\theta_0 + \frac{M}{\sqrt{n}}} Q^n_{\theta',\zeta}\biggl|\sum_{i=1}^n g_{\theta',\zeta}(X_i)\biggr|\, d\theta' \leq \sqrt{n} \int_{\theta_0 - \frac{M}{\sqrt{n}}}^{\theta_0 + \frac{M}{\sqrt{n}}} \sqrt{Q_{\theta',\zeta} g^2_{\theta',\zeta}}\, d\theta', \end{aligned}$$

where the last step follows from the Cauchy-Schwartz inequality. For large enough n, $\rho_n < \rho$ and the square-root of (17.27) dominates the difference between $U(\rho, h_n)$ and 1. □

17.3 Posterior asymptotic normality

Under the assumptions formulated before theorem 15.7, the marginal posterior density $\pi_n(\cdot|X_1, \ldots, X_n) \colon \Theta \to \mathbb{R}$ for the parameter of interest with respect to the prior Π_Θ equals,

$$\pi_n(\theta|X_1, \ldots, X_n) = S_n(\theta) \;/\; \textstyle\int_\Theta S_n(\theta')\, d\Pi_\Theta(\theta'), \tag{17.28}$$

P_0^n-almost-surely. One notes that this form is equal to that of a *parametric* posterior density, but with the parametric likelihood replaced by the integrated likelihood S_n. By implication, the proof of the parametric Bernstein-Von Mises theorem can be applied to its semiparametric generalization, if we impose sufficient conditions for the parametric likelihood on S_n instead. Concretely, we replace the smoothness requirement for the likelihood in theorem 14.3 by (17.16). Together with a condition expressing marginal posterior convergence at parametric rate, (17.16) is sufficient to derive asymptotic normality of the posterior *c.f.* (15.3).

Theorem 17.8 (Posterior asymptotic normality)
Let Θ be open in $\mathbb{R}^k$ with a prior Π_Θ that is thick at θ_0. Suppose that for large enough n,

the map $h \mapsto s_n(h)$ is continuous P_0^n-almost-surely. Assume that there exists an $L_2(P_0)$-function $\tilde{\ell}_{\theta_0,\eta_0}$ such that for every (h_n) that is bounded in probability, (17.16) holds, $P_0\tilde{\ell}_{\theta_0,\eta_0} = 0$ and $\tilde{I}_{\theta_0,\eta_0}$ is non-singular. Furthermore suppose that for every (M_n), $M_n \to \infty$, we have:

$$\Pi_n\big(\, \|h\| \leq M_n \;\big|\; X_1, \ldots, X_n \,\big) \xrightarrow{P_0} 1. \tag{17.29}$$

Then the sequence of marginal posteriors for θ converges to a normal distribution in total variation,

$$\sup_A \Big| \Pi_n\big(\, h \in A \;\big|\; X_1, \ldots, X_n \,\big) - N_{\tilde{\Delta}_n, \tilde{I}^{-1}_{\theta_0,\eta_0}}(A) \Big| \xrightarrow{P_0} 0,$$

centred on $\tilde{\Delta}_n$ with covariance matrix $\tilde{I}^{-1}_{\theta_0,\eta_0}$.

Proof Throughout we denote the normal distribution centred on $\tilde{\Delta}_n$ with covariance $\tilde{I}^{-1}_{\theta_0,\eta_0}$ by Φ_n. The prior and marginal posterior for the local parameter h are denoted Π_n and $\Pi_n(\,\cdot\,|X_1, \ldots, X_n)$. Conditioned on some C measurable in $\mathbb{R}^k$, we denote these measures by Φ_n^C, Π_n^C and $\Pi_n^C(\,\cdot\,|X_1, \ldots, X_n)$ respectively.

Let C be compact in $\mathbb{R}^k$ and assume that C contains an open neighbourhood of the origin. Define, for every $g, h \in C$ and large enough n,

$$f_n(g,h) = \left(1 - \frac{\phi_n(h)}{\phi_n(g)} \frac{s_n(g)}{s_n(h)} \frac{\pi_n(g)}{\pi_n(h)} \right)_+,$$

where $\phi_n\colon C \to \mathbb{R}$ is the Lebesgue density of the distribution Φ_n and $\pi_n\colon C \to \mathbb{R}$ is the Lebesgue density of the prior Π_n for the parameter h. By assumption (17.16) we have, for every stochastic (h_n) in C:

$$\begin{aligned} \log s_n(h_n) &= \log s_n(0) + h_n^T \mathbb{G}_n \tilde{\ell}_{\theta_0,\eta_0} - \tfrac{1}{2} h_n^T \tilde{I}_{\theta_0,\eta_0} h_n + o_{P_0}(1), \\ \log \phi_n(h_n) &= -\tfrac{1}{2}(h_n - \tilde{\Delta}_n)^T \tilde{I}_{\theta_0,\eta_0} (h_n - \tilde{\Delta}_n) + D_n, \end{aligned}$$

(with normalization constants D_n that cancel in the fraction that defines f_n). For any two stochastic sequences (h_n), (g_n) in C, $\pi_n(g_n)/\pi_n(h_n)$ converges to 1 as $n \to \infty$. Combining with the above display and with (15.4), we see that:

$$\begin{aligned} \log \frac{\phi_n(h_n)}{\phi_n(g_n)} \frac{s_n(g_n)}{s_n(h_n)} \frac{\pi_n(g_n)}{\pi_n(h_n)} &= -h_n^T \mathbb{G}_n \tilde{\ell}_{\theta_0,\eta_0} + \tfrac{1}{2} h_n^T \tilde{I}_{\theta_0,\eta_0} h_n + g_n^T \mathbb{G}_n \tilde{\ell}_{\theta_0,\eta_0} - \tfrac{1}{2} g_n^T \tilde{I}_{\theta_0,\eta_0} g_n + o_{P_0}(1) \\ &\quad - \tfrac{1}{2}(h_n - \tilde{\Delta}_n)^T \tilde{I}_{\theta_0,\eta_0} (h_n - \tilde{\Delta}_n) + \tfrac{1}{2}(g_n - \tilde{\Delta}_n)^T \tilde{I}_{\theta_0,\eta_0} (g_n - \tilde{\Delta}_n) \\ &= o_{P_0}(1), \end{aligned} \tag{17.30}$$

as $n \to \infty$. For any stochastic sequence (h_n, g_n) in $C \times C$, $f_n(g_n, h_n) \xrightarrow{P_0} 0$, by continuous mapping. By (17.16), $s_n(h)/s_n(0)$ is of the form $\exp(K_n(h) + R_n(h))$ for all h and $n \geq 1$, where $R_n = o_{P_0}(1)$. Tightness of K_n and R_n implies that $s_n(h)/s_n(0) \in (0, \infty)$, P_0^n-almost-surely. Almost-sure continuity of $h \mapsto s_n(h)$ then implies almost-sure continuity of $(g,h) \mapsto s_n(g)/s_n(h)$ for large enough n. Since $\tilde{\ell}_{\theta_0,\eta_0} \in L_2(P_0)$ and $\tilde{I}_{\theta_0,\eta_0}$ is invertible, the location of the normal distribution $N_{\tilde{\Delta}_n, \tilde{I}_0}$ is P_0^n-tight, so that $(g,h) \mapsto \phi_n(g)/\phi_n(h)$ is continuous on $C \times C$. The thickness of the prior density π guarantee that this also holds for

$(g,h) \mapsto \pi_n(g)/\pi_n(h)$. Since, for large enough n, f_n is continuous on $C \times C$, P_0^n-almost-surely, we conclude that the convergence of f_n holds uniformly over $C \times C$, *i.e.*,

$$\sup_{g,h \in C} f_n(g,h) \xrightarrow{P_0} 0. \tag{17.31}$$

For given $\delta > 0$, define the events $\Omega_n = \{\sup_{g,h\in C} f_n(g,h) \le \delta\}$, so that, Because C contains a neighbourhood of the origin and $\tilde{\Delta}_n$ is tight for all $n \ge 1$, $\Phi_n(C) > 0$, P_0^n-almost-surely. Moreover, the prior mass of C satisfies $\Pi_n(C) > 0$ and for all $h \in C$, $s_n(h) > 0$, so that the posterior mass of C satisfies $\Pi_n(C|X_1,\ldots,X_n) > 0$. Therefore, conditioning on C is well-defined P_0^n-almost-surely for both Φ_n and $\Pi_n(\,\cdot\,|X_1,\ldots,X_n)$. We consider the difference in total variation between $\Pi_n^C(\,\cdot\,|X_1,\ldots,X_n)$ and Φ_n^C. We decompose its P_0^n-expectation and use (17.31) to conclude that,

$$\begin{aligned} &P_0^n \sup_A \bigl|\Pi_n^C(A|X_1,\ldots,X_n) - \Phi_n^C(A)\bigr| \\ &\qquad\qquad \le P_0^n \sup_A \bigl|\Pi_n^C(A|X_1,\ldots,X_n) - \Phi_n^C(A)\bigr| 1_{\Omega_n} + o_{P_0}(1). \end{aligned} \tag{17.32}$$

Note that both Φ_n^C and $\Pi_n^C(\,\cdot\,|X_1,\ldots,X_n)$ have strictly positive densities on C for large enough n. Therefore, Φ_n^C is dominated by $\Pi_n^C(\,\cdot\,|X_1,\ldots,X_n)$ if n is large enough. The former term on the *r.h.s.* in (17.32) can now be calculated as follows:

$$\begin{aligned} &\tfrac{1}{2} P_0^n \sup_A \bigl|\Pi_n^C(A|X_1,\ldots,X_n) - \Phi_n^C(A)\bigr| 1_{\Omega_n} \\ &\quad = P_0^n \int_C \Bigl(1 - \int_C \frac{s_n(g)\pi_n(g)\phi_n(h)}{s_n(h)\pi_n(h)\phi_n(g)} d\Phi_n^C(g)\Bigr)_+ d\Pi_n^C(h|X_1,\ldots,X_n)\, 1_{\Omega_n}, \end{aligned}$$

for large enough n. Jensen's inequality and substitution of (17.31) then gives,

$$\begin{aligned} &\tfrac{1}{2} P_0^n \sup_{A\in\mathscr{B}} \bigl|\Pi_n^C(A|X_1,\ldots,X_n) - \Phi_n^C(A)\bigr| 1_{\Omega_n} \\ &\qquad \le P_0^n \int \sup_{g,h\in C} f_n(g,h)\, 1_{\Omega_n}\, d\Phi_n^C(g)\, d\Pi_n^C(h|X_1,\ldots,X_n) \le \delta, \end{aligned}$$

for large enough n. Since the argument holds for all $\delta > 0$, substitution of (17.32) shows that for all compact $C \subset \mathbb{R}^k$ containing a neighbourhood of 0,

$$P_0^n \bigl\|\Pi_n^C - \Phi_n^C\bigr\| \to 0.$$

Let (B_m) be a sequence of closed balls centred at the origin with radii $M_m \to \infty$. For each fixed $m \ge 1$, the above display holds with $C = B_m$, so if we choose a sequence of balls (B_n) that traverses the sequence (B_m) slowly enough, convergence to zero can still be guaranteed. We conclude that there exists a sequence of radii (M_n) such that $M_n \to \infty$ and,

$$P_0^n \bigl\|\Pi_n^{B_n} - \Phi_n^{B_n}\bigr\| \to 0. \tag{17.33}$$

Combining (17.29) and lemma 5.2 in Kleijn and van der Vaart (2007) (the sequence $(\tilde{\Delta}_n)$ converges weakly, so that it is uniformly tight by Prohorov's theorem), we then use lemma 5.1 in Kleijn and van der Vaart (2007) to conclude that

$$P_0^n \bigl\|\Pi_n - \Phi_n\bigr\| \to 0,$$

the assertion holds. □